Language Learning Tasks and Automatic Analysis of Learner Language:

Connecting FLTL and NLP design of ICALL materials

supporting use in real-life instruction

Dissertation
zur
Erlangung des akademischen Grades
Doktor der Philosophie
in der Philosophischen Fakultät

der Eberhard Karls Universität Tübingen

vorgelegt von

Martí Quixal Martínez

aus

Sant Boi de Llobregat, Catalunya (Spanien)

2015

Gedruckt mit Genehmigung der Philosophischen Fakultät
der Eberhard Karls Universität Tübingen

Dekan: Prof. Dr. Jürgen Leonhardt

Hauptberichterstatter: Prof. Dr. Walt Detmar Meurers
Mitberichterstatter: Dr. Toni Badia i Cardús (Universitat Pompeu Fabra in Barcelona)
Tag der mündlichen Prüfung: 20. Dezember, 2012

Universitätsbibliothek Tübingen, TOBIAS-lib

Language Learning Tasks and Automatic Analysis of Learner Language

Connecting FLTL and NLP in the design of ICALL materials supporting effective use in real-life instruction

Martí Quixal Martínez

DOCTORAL THESIS

CO-ADVISORS

Dr. Toni Badia
Departament de Traducció i Ciències del Llenguatge
Universitat Pompeu Fabra

Prof. Dr. Walt Detmar Meurers
Seminar für Sprachwissenschaft
Eberhard-Karls-Universität Tübingen

Universitat Pompeu Fabra Barcelona

Eberhard Karls Universität Tübingen

Als meus pares.

A l'Angelina i a la Bruna.

Jo no tinc cançons;
em tenen a mi
elles, les cançons.
Quan volen, quan vénen,
quan? Qui ho pot saber.
(...)
Jo he passat hores, dies i anys
per cases, per carrers i per ciutats,
per boscos i camins, per vents i mars
percaçant-les. Oh, desig de cançons.
(...)

Oh, desig de cançons
Raimon

Agraïments

Aquesta tesi hagués estat del tot impossible sense l'ajuda, la col·laboració i la paciència de moltes persones. Totes elles m'han ajudat a fer-la una mica millor, i només jo sóc responsable de les imperfeccions que hi restin.

Vull donar les gràcies molt especialment als meus directors de tesi: el Dr. Toni Badia i Cardús, i el Prof. Dr. Walt Detmar Meurers. Sense ells aquesta tesi no s'hagués dut mai a terme. Al Dr. Badia he d'agrair-li que m'acceptés com a estudiant de doctorat el 1999 i que tingués la virtut de deixar-me fer les coses a la meva manera. L'autonomia que sempre m'ha donat, acompanyada d'excel·lents consells acadèmics quan aquests eren necessaris o sol·licitats, han fet que el camí recorregut hagi sigut instructiu, enriquidor i ple de reptes professionals i acadèmics convertits ara en unes experiències i una satisfacció impagables.

Herrn Prof. Dr. Meurers bin ich zu unendlichem Dank verpflichtet, er half mir diese Dissertation zu Ende zu bringen, als ich nicht mehr die Kraft dazu hatte. Wissenschaftlich lehrte er mich Exzellenz und Konsequenz, vor allem ermöglichte er es mir, in einem der für mich interessantesten wissenschaftlichen Themen zu forschen – mit festem Schritt, mit Freude und mit Freundschaft. Ihm danke ich auch dafür, mir einen fünfmonatigen Aufenthalt am Seminar für Sprachwissenschaft der Eberhard Karls Universität Tübingen ermöglicht zu haben.

I also want to thank Dr. Pérez-Vidal and Dr. Boleda for reading earlier versions of some of the chapters of this dissertation and having provided me with very valuable scientific and academic feedback.

There are many other people whom I am grateful to. Among them, there are the researchers, the technicians and the practitioners with whom I collaborated in the past ten years, particularly during the three projects that provided me with a large research and experimental basis to carry out my research: the ALLES, the AutoLearn and the ICE^3 projects. All of them appear to me as indispensable today.

However, there are some people with whom I worked more intensively, and to these people I want to give special thanks.

Vull agrair al Francesc Benavent, l'Stefan Bott, el Beto Boullosa, la Mariona Estrada, el David García Narbona, l'Àngel Gil, l'Araceli Martínez i l'Oriol Valentín l'esforç, la intel·ligència i la il·lusió que van aportar a totes les tasques que conjuntament vam dur a terme al llarg de diversos projectes. Sense ells, aquesta tesi hauria estat avorrida i impossible.

Quiero agradecer a la Dra. Lourdes Díaz, Ana Ruggia y Rosa Lucha el intenso e interesante trabajo que pudimos realizar conjuntamente durante el proyecto ALLES. Sin su colaboración, me hubiera sido imposible adentrarme en y comprender el mundo de la didáctica de lenguas.

Den Mitgliedern des Instituts der Gesellschaft zur Förderung der Angewandten Informationsforschung an der Universität des Saarlandes möchte ich dafür danken, dass sie immer dazu bereit waren, ihr Wissen, ihre Erfahrung und ihre Tools mit mir zu teilen. Dies geschah mit Professionalität und Wärme. Danke Dagmar, Prof. Dr. Haller, Dr. Preuß, Dr. Schmidt, Sandrine, Ute und Volkan.

I am also thankful to my colleagues in Universidad Europea de Madrid, Heriot-Watt University, in Boğaziçi University and in TEHNE. They all provided a stimulating working environment.

Unes persones a qui mai no podré agrair prou la seva col·laboració són els i les mestres i els i les estudiants que han participat en els diversos experiments que hem fet tant en contextos de laboratori com en contextos d'instrucció real. Entre aquestes persones vull esmentar l'Anna Campillo, la Mònica Castanyer, el Mikel Martín, la Montse Paradeda, la Gemma Pou i tots els seus i les seves alumnes. Aquests cinc mestres i les seves classes m'han demostrat que en el meu país hi ha gent disposada a arriscar-se, i a fer-ho per convicció i amb l'objectiu d'aprendre. Tots ells, totes elles, són els que donen sentit a aquesta tesi.

També han estat molt importants les persones que treballen a la secretaria del Departament de Traducció i Ciències del Llenguatge. Vull donar les gràcies especialment a la Susi Bolós i la Lali Palet, que sempre m'han ajudat amb totes les gestions que he hagut de fer amb gran professionalitat i prestesa.

En un terreny més personal vull donar les gràcies als meus pares, Sebastià i Angelina, per haver-me dut al món i per haver-me ensenyat a gaudir de les coses bones, entre les quals hi compta la il·lusió per fer una feina que t'agradi i el compromís de fer-la honestament el millor possible. També han estat tot amor i ajuda al llarg de la meva vida.

Altres membres de família, de sang i polítics, també han fet més fàcil aquesta tesi, tant en aspects pràctics a l'hora de facilitar-me la feina, com en aspectes personals a l'hora d'escoltar i donar ànims: Sebastià, Miquel, tia Montse, Rebeca, Esther, Margarita, Josep Maria, a tots vosaltres us estic molt agraït.

També vull donar les gràcies a les meves filles, la Bruna i l'Angelina, que amb el seu amor incondicional i la seva curiositat innata, i, de vegades, esgotadora, m'han anat recordant que cal esforçar-se per fer el que t'agrada, i el que et convenç. Sempre han suportat els mals dies i, a més, m'han animat molt: pa-pa, pa-pa!

I, finalment, una persona que va ser important i molt fins a la consecució i la defensa d'aquesta tesi, la Gemma Boleda. Ella va patir aquesta tesi durant deu anys: a l'hora de fer el cafè, a l'hora de menjar i a l'hora de dormir. Se la va endur amb mi al cine, al metge, de vacances, als congressos, als cursos d'estiu, de bodes, dues vegades a la sala de parts, i també a algun enterrament. Aquesta tesi fou passejada fins a l'extenuació. Sempre li n'estaré agraït.

Barcelona, 9 de juliol del 2013

Funding

This thesis was possible to a large extent thanks to the following funding institutions or projects:

- Universitat Pompeu Fabra, from which I received a pre-doctoral research fellowship from December 1999 to Dec 2001 to work at the Institut Universitari de Lingüística Aplicada;
- ALLES project, funded by the European Commission under the 5th Framework Programme, contract number IST-2001-34246:
- AutoLearn project, funded by the Education, Audiovisual and Culture Executive Agency under the Lifelong Learning Programme, project number 135693-LLP-1-2007-1-ES-KA3-KA3MP:
- ICE^3 project, funded by the Education, Audiovisual and Culture Executive Agency under the Lifelong Learning Programme, project number 510653-LLP-1-2010-1-ES-COMENIUS-CMP:
- Ministerio de Educación del Gobierno de España, Subvención para la movilidad de estudiantes para la obtenició del Doctorado Europeo, proyecto número TME2009-00266; and,
- Fundació Barcelona Media, whose Executive Committee conceded me an 11-month sabbatical period for me to work exclusively on my dissertation.

Abstract

This thesis studies the application of Natural Language Processing to Foreign Language Teaching and Learning, within the research area of Intelligent Computer-Assisted Language Learning (ICALL). In particular, we investigate the design, the implementation, and the use of ICALL materials to provide learners of foreign languages, particularly English, with automated feedback.

We argue that the successful integration of ICALL materials demands a design process considering both pedagogical and computational requirements as equally important. Our investigation pursues two goals. The first one is to integrate into task design insights from Second Language Acquisition and Foreign Language Teaching and Learning with insights from computational linguistic modelling. The second goal is to facilitate the integration of ICALL materials in real-world instruction settings, as opposed to research or lab-oriented instruction settings, by empowering teachers with the methodology and the technology to autonomously author such materials.

To achieve the first goal, we propose an ICALL material design process that combines basic principles of Task-Based Language Instruction and Task-Based Test Design with the specification requirements of Natural Language Processing. The relation between pedagogical and computational requirements is elucidated by exploring (i) the formal features of foreign language learning activities, (ii) the complexity and variability of learner language, and (iii) the feasibility of applying computational techniques for the automatic analysis and evaluation of learner responses.

To achieve the second goal, we propose an automatic feedback generation strategy that enables teachers to customise the computational resources required to automatically correct ICALL activities without the need for programming skills. This proposal is instantiated and evaluated in real world-instruction settings involving teachers and learners in secondary education.

Our work contributes methodologically and empirically to the ICALL field, with a novel approach to the design of materials that highlights the cross-disciplinary and iterative nature of the task. Our findings reveal the strength of characterising tasks both from the perspective of Foreign Language Teaching and Learning and from the perspective of Computational Linguistics as a means to clarify the nature of learning activities. Such a characterisation allows us to identify ICALL materials which are both pedagogically meaningful and computationally feasible.

Our results show that teachers can characterise, author and employ ICALL materials as part of their instruction programme, and that the underlying computational machinery can provide the required automatic processing with sufficient efficiency. The authoring tool and the accompanying methodology become a crucial instrument for ICALL research and practice: Teachers are able to design activities for their students to carry out without relying on an expert in Natural Language Processing. Last but not least, our results show that teachers are value the experience very positively as means to engage in technology integration, but also as a means to better apprehend the nature of their instruction task. Moreover, our results show that learners are motivated by the opportunity of using a technology that enhances their learning experience.

Resum

Aquest treball de recerca es troba a la cruïlla entre el Processament del Llenguatge Natural i l'Aprenentatge i Ensenyament de Llengües Estrangeres i, en concret, dins l'àrea anomenada Aprenentatge de Llengües Assistit per Ordinador amb Intel·ligència Artificial (en anglès, *Intelligent Computer-Assisted Language Learning*, abreujat ICALL). La nostra recerca se centra en el disseny, la implemenació i l'ús de materials d'ICALL per proveir els estudiants de llengües estrangeres, i especialment d'anglès, de materials que incorporin funcionalitats de correcció i avaluació automàtiques de les respostes.

En aquesta tesi defensem que, per tal que la integració de materials d'ICALL sigui reeixida, cal tenir en compte per igual els requisits pedagògics i els computacionals ja en la fase de disseny dels materials. Els nostres objectius principals són dos. D'una banda, volem integrar en el procés de disseny de materials tant els principis fonamentals de l'Adquisició de Segones Llengües i l'Aprenentatge i Ensenyament de Llengües Estrangeres com els principis fonamentals del modelatge lingüístic. D'altra banda, volem facilitar la integració dels materials d'ICALL en contextos d'instrucció reals, en contrast amb els contextos d'instrucció de recerca o de laboratori, per tal de capacitar els docents amb la metodologia i la tecnologia necessàries perquè puguin crear autònomament materials d'ICALL.

Per aconseguir el primer objectiu, proposem un procés de disseny de materials d'ICALL que combina els principis bàsics de l'Ensenyament de Llengües Basat en Tasques i el Disseny de Tests Basat en Tasques amb la mena d'especificacions requerides per les eines de Processament del Llenguatge Natural. Explorem la relació entre els requisits pedagògics i computacionals des de tres punts de vista: (i) les característiques formals de les activitats per a l'aprenentatge de llengües estrangeres, (ii) la complexitat i la variabilitat de la llengua dels estudiants, i (iii) la viabilitat d'aplicar tècniques computacionals per a l'anàlisi i avaluació automàtiques de les respostes.

Per aconseguir el segon objectiu, proposem una estratègia d'avaluació automàtica que permet als i les docents adaptar els recursos lingüístics computacionals necessaris per corregir automàticament les activitats d'ICALL sense la necessitat d'aprendre de programar. Per provar la viabilitat de la proposta presentem un experiment en què l'apliquem i l'avaluem en entorns d'aprenentatge reals amb docents i aprenents d'educació secundària.

Amb aquesta tesi fem una contribució metodològica i empírica al camp de l'ICALL, amb una aproximació innovadora al disseny de materials que posa èmfasi en la naturalesa multidisciplinar i iterativa del procés. Els resultats que presentem revelen el potencial de la caracterització de tasques d'aprenentatge conjugant la perspectiva l'Aprenentatge i Ensenyament de Llengües Estrangeres i la de la Lingüísitica Computacional com un instrument clau per descriure formalment les activitats d'aprenentatge. Aquesta caracterització permet identificar materials d'ICALL que siguin alhora pedagògicament rellevants i computacionalment viables.

Els resultats demostren que amb l'estratègia proposada els i les docents poden caracteritzar, crear i emprar materials d'ICALL dins del seu programa d'instrucció, i que el programari computacional subjacent proporciona el processament automàtic

requerit amb una qualitat acceptable per a l'ús en contextos d'instrucció reals. El programari i la metodologia proposats esdevenen crucials per a la recerca i la pràctica de l'ICALL: els docents són capaços de dissenyar activitats per als estudiants sense dependre d'un expert en Processament del Llenguatge Natural. Finalment, els resultats també demostren que els i les docents valoren l'experiència molt positivament en la mesura que els permet integrar noves tecnologies a l'aula, i alhora els permet comprendre millor la naturalesa de la seva tasca docent. A més, els resultats demostren que els i les estudiants se senten motivats pel fet de poder emprar una tecnologia que permet una avaluació immediata i personalitzada de la seva activitat d'aprenentatge.

Zusammenfassung

Diese Dissertation thematisiert die Schnittstelle zwischen maschineller Sprachverarbeitung und der Fremdsprachenlehre. Die Untersuchung ist Teil des Forschungsbereichs des Computerunterstützten Sprachenlernens mithilfe Künstlicher Intelligenz (Intelligent Computer-Assisted Language Learning, ICALL). Im Besonderen untersucht werden sollen die Gestaltung, die Implementierung und der Einsatz von ICALL-Materialien, die dem Lernenden von Fremdsprachen, insbesondere dem Englischen, automatisches Feedback liefern sollen.

Wir zeigen, dass ein erfolgreicher Einsatz von ICALL-Materialien einen Designprozess verlangt, der pädagogische und computerlinguistische Anforderungen gleichermaßen berücksichtigt. Unsere Untersuchung verfolgt zwei Ziele. Erstens sollen Erkenntnisse aus dem Fremdspracherwerb, dem Fremdsprachenlehren und -lernen und Erkenntnisse aus der Computerlinguistik beim Erzeugen von Lernaufgaben kombiniert werden. Zweitens soll die Anwendung von ICALL-Materialien in realen Unterrichtssituationen, im Gegensatz zu experimentellen oder Forschungs-Unterrichtssituationen, ermöglicht werden. Hierbei soll den Lehrenden die entsprechende Methodik und Technologie vermittelt werden, so dass sie in die Lage versetzt werden, selbständig solche Materialen zu gestalten.

Zur Erreichung des ersten Zieles wird ein Designprozess von ICALL-Materialien vorgeschlagen, der Grundprinzipien des aufgabenbasierten Fremdsprachenerwerbes und -unterrichtes (im Sinne von Task-Based Language Instruction) sowie aufgabenbasiertes Sprachtestdesign mit den Spezifikationsanforderungen der maschinellen Sprachverarbeitung verbindet. Die Beziehung zwischen pädagogischen und computerlinguistischen Anforderungen wird untersucht durch eine Analyse i) der Merkmale der Lernaufgabe, ii) der Komplexität und Variabilität der Sprache des Lernenden und iii) der Realisierbarkeit der computerlinguistischen Techniken für die automatische Analyse und Bewertung der Lernerantworten.

Der Erreichung des zweiten Zieles dient die Generierung eines automatischen Feedbackprozesses, der es Lehrern ermöglichen soll, die Datenquellen, die für die automatische Korrektur notwendig, sind entsprechend anzupassen, ohne dass sie dazu Programmierkenntnisse benötigen. Dieses Unterfangen wird beispielhaft anhand von realen Unterrichtssituationen mit Lehrern und Lernern der Sekundärstufe untersucht und bewertet.

Diese Dissertation liefert einen methodischen und empirischen Beitrag zum ICALL Forschungsbereich dar. Sie charakterisiert einen neuen Ansatz in Bezug auf die Gestaltung von Lernmaterial und betont dabei die Interdisziplinarität und die iterative Natur dieser Aufgabe. Die Ergebnisse dieser Arbeit zeigen die Stärke einer Methode der Charakterisierung von Lernaufgaben aus zwei Perspektiven auf: die der Fremdsprachlehre und -didaktik und die der Computerlinguistik. Nur die Kombination beider Perspektiven ermöglicht die Erstellung von ICALL-Materialien die sowohl pädagogisch sinnvoll als auch computerlinguistisch realisierbar sind.

Die Forschungsergebnisse zeigen, dass Lehrer ICALL-Materialien als Teil ihres Unterrichtsprogramms verfassen und anwenden können, und dass die zugrundeliegende computerlinguistische Verarbeitung die erforderlichen automatischen Prozesse effizient leisten kann. Das Authoring-System mithilfe dessen die Lehrer Aufgaben

verfassen können sowie die begleitende Methodik werden zu entscheidenden Instrumenten der ICALL Forschung und Praxis. Lehrer können damit Aufgaben für ihre Schüler gestalten ohne dazu einen computerlinguistischen oder informatischen Experten zu benötigen. Die Evaluation dieses Ansatzes zeigt dabei auch, dass Lehrer ihre Erfahrung mit dem System äußerst positiv bewerten, da das System zum einen einen sinnvollen Einsatz von aktueller Sprachverarbeitungstechnologie ermöglicht, aber zum anderen auch eine Möglichkeit bietet, die Charakteristika der Aufgaben und ihrer Lehrtätigkeit besser zu verstehen. Die Forschungsergebnisse zeigen zudem, dass der Einsatz der ICALL Technologie den Lernprozess auch für die Lerner attraktiver macht.

Contents

List of Figures xxxiii

List of Tables xxxviii

I Introduction 1

1 Motivation and research goals 5
- 1.1 Computers in foreign language learning 6
- 1.2 NLP as a transferable technology 7
- 1.3 The interrelationship between NLP and FLTL 9
- 1.4 ICALL in real-world instruction settings 9
- 1.5 Research goals 10
- 1.6 Structure of the thesis 11

2 The goals of the thesis within ICALL 15
- 2.1 An overview of the research in ICALL 15
  - 2.1.1 The beginning of ICALL 15
  - 2.1.2 More than 30 years of ICALL 17
  - 2.1.3 The essence of an ILTS 19
    - 2.1.3.1 Architecture and functionalities of an ILTS 20
  - 2.1.4 ICALL systems in use 21
- 2.2 Task design and automatic language processing 26
  - 2.2.1 The pedagogical purpose as a driver of ICALL research 26
  - 2.2.2 From the focus on form to the focus on meaning 27
- 2.3 Tools for teachers to author FL activities 29
  - 2.3.1 Teacher control over CALL materials 29
  - 2.3.2 Tutor Assistant 30
  - 2.3.3 Automatic generation of ICALL activities 31
- 2.4 Revisiting the goals of this thesis 31
  - 2.4.1 The feasibility of ICALL 32
  - 2.4.2 An autonomous use of ICALL in class 33
- 2.5 Chapter summary 34

II Background 35

3 Natural Language Processing 39
3.1 Fundamental concepts in NLP . . . 39
3.1.1 Approaches to processing natural language . . . 40
3.1.1.1 Deep versus shallow NLP processing . . . 41
3.1.2 The domain of application . . . 42
3.1.2.1 The domain in foreign language teaching and learning 43
3.1.3 Robust NLP tools . . . 44
3.2 Analysing learner language . . . 45
3.2.1 Symbolic approaches to process ill-formed language . . . 45
3.2.1.1 Mal-rule approach . . . 45
3.2.1.2 Constraint relaxation approach . . . 46
3.2.1.3 Pros and cons . . . 47
3.2.2 Stochastic approaches to detect deviations from the norm . . . 48
3.3 Chapter summary . . . 49

4 Foreign Language Teaching and Learning 51
4.1 Modern instruction of foreign languages . . . 52
4.2 TBLT: principles and practice . . . 53
4.2.1 Task-Based Language Instruction . . . 53
4.2.1.1 Analysing the properties of language learning activities 54
4.2.2 The design of a TBLT syllabus . . . 55
4.2.2.1 The linguistic contents of tasks . . . 58
4.3 Assessment of learner production . . . 60
4.3.1 Summative assessment . . . 61
4.3.1.1 A framework for the characterisation of test tasks . . 61
4.3.2 Formative assessment . . . 63
4.4 Feedback as a means to help learners . . . 63
4.4.1 Types of feedback . . . 64
4.4.2 The effectiveness of feedback . . . 64
4.5 Feedback studies in CALL . . . 65
4.5.1 The effectiveness of feedback in CALL . . . 65
4.5.2 Computer-based feedback vs. teacher feedback . . . 66
4.5.3 The use of CALL feedback by learners . . . 66
4.5.4 The use of ICALL feedback by learners . . . 67
4.6 Chapter summary . . . 68

III ICALL tasks – Where FLTL meets NLP 71

5 Methodological considerations 75
5.1 Teaching and learning in an ICALL setting . . . 75
5.1.1 Interaction flow in an ICALL setting . . . 75
5.2 The life cycle of ICALL tasks . . . 76
5.2.1 Interaction flow in the execution phase . . . 78

5.2.2 Interaction flow in the design phase 78
5.2.3 Interrelationships in the evaluation phase 79
5.3 Connecting FLTL and NLP in the lifecycle of ICALL tasks 80
5.4 Chapter summary 81

6 A research setting to develop ICALL materials 83
6.1 Overall perspective 83
6.2 TBLT-driven design of materials 85
6.2.1 Determining an interest area 85
6.2.2 Planning a final task 86
6.2.3 Determine the unit objectives 87
6.2.4 Content specification of the unit of work 87
6.2.5 Process plan 88
6.2.5.1 Learning sequences in ALLES 89
6.2.6 Instruments and procedure for evaluation 90
6.2.6.1 Formative assessment in ALLES 90
6.2.6.2 Summative assessment in ALLES 91
6.2.7 From the design to the actual materials 91
6.3 A general architecture for the analysis of learner language 92
6.3.1 The linguistic analysis underlying domain-specific assessment . 94
6.3.2 Two concrete implementations 95
6.3.2.1 The MPRO-KURD solution 96
6.3.2.2 The CG-based solution 99
6.3.3 KURD and CG for shallow semantic processing 102
6.3.3.1 CG-based shallow semantic processing 102
6.3.3.2 KURD-based shallow semantic processing 103
6.4 Chapter summary 104

7 Designing ICALL tasks – Characterisation of pedagogical needs 105
7.1 TAF: Task Analysis Framework 105
7.1.1 Definition 105
7.1.2 Applying the TAF to *Education and Training* 107
7.1.2.1 Introduction and pre-test 107
7.1.2.2 “Having a well-motivated workforce” 108
7.1.2.3 “Recommend a course and ask for information” . . . 110
7.1.2.4 “Asking information about a course” 112
7.1.2.5 “Registering for a course” 114
7.1.2.6 *Education and Training* as a whole 116
7.1.3 FL learning tasks as candidates to become ICALL tasks . . . 116
7.2 RIF: the Response Interpretation Framework 117
7.2.1 Definition 117
7.2.2 Applying the RIF to four FL learning tasks 119
7.2.2.1 Task type I 119
7.2.2.2 Task type II 124
7.2.2.3 Task type III 131
7.2.2.4 Task type IV 145

7.3 Chapter summary 150

8 NLP functionalities to respond to FLTL demands 153

8.1 From pedagogical requirements to specifications for the Automatic Assessment 154

8.1.1 AASF: Automatic Assessment Specification Framework 154

8.1.1.1 SALA: Specifications for Automatic Linguistic Analysis 154

8.1.1.2 SFGL: Specifications for the Feedback Generation Logic 155

8.2 Applying the AASF to ICALL tasks 156

8.2.1 Applying the AASF for formative assessmemt 156

8.2.1.1 The SALA applied to formative feedback 156

8.2.1.2 The SFGL applied to an activity with formative feedback 160

8.2.2 Applying the AASF for summative assessment 165

8.2.2.1 The SALA applied to a task for summative feedback 165

8.2.2.2 The SFGL applied to an activity with summative feedback 170

8.3 A feedback generation strategy for the assessment of ICALL activities 173

8.3.1 A general NLP-based architecture for the automatic assessment of learner responses 174

8.3.2 The point of departure for NLP-based automatic assessment . 174

8.4 Automatic generation of formative feedback 176

8.4.1 Modelling automatic assessment for correct responses 176

8.4.1.1 Modelling linguistic analysis of correct responses . . 176

8.4.1.2 Modelling feedback generation of correct responses . 178

8.4.2 Modelling incorrect responses 178

8.4.2.1 Modelling wrong choices in responses 179

8.4.2.2 Modelling missing or unexpected information 184

8.4.3 Modelling extended production responses 187

8.4.4 Modelling loosely restricted production responses 190

8.5 Automatic generation of summative feedback 190

8.5.1 Analysing learner responses for the generation of summative feedback 191

8.5.2 Evaluating learner responses for the generation of summative feedback 192

8.6 Chapter summary 193

9 ICALL task complexity on the basis of learner data 195

9.1 Annotation of learner responses 196

9.1.1 Comparing design-based specifications and learner responses . 196

9.1.1.1 Correctness and well-formedness 197

9.1.2 Scheme for the annotation of learner responses 198

9.2 Learner language in task responses 199

9.2.1 Responses to a Type I task 199

9.2.1.1 Qualitative analysis of the language of learner responses 200

9.2.2 Responses to a Type III activity 206

9.2.2.1 Qualitative analysis of the language of learner responses 206
9.2.3 Responses to a Type IV task 213
9.2.3.1 Qualitative analysis of the language of learner responses 213
9.3 Response characteristics and NLP complexity 220
9.3.1 Response length 220
9.3.2 Response variation 221
9.3.3 Section summary 223
9.4 Learner data to improve the analysis strategy 223
9.4.1 Corpus-driven domain adaptation 223
9.4.2 Corpus-driven mal-rule approach 224
9.5 Chapter summary 224

IV Enabling teachers to author ICALL activities 227

10 Customisation of an NLP-based feedback generation strategy 231
10.1 Context of application 232
10.1.1 Formative feedback as a functionality 233
10.2 Customisable NLP-feedback assessment 234
10.2.1 A customisable architecture 234
10.2.2 Response Specification Language 235
10.2.2.1 Definition of the RSL 236
10.2.2.2 RSL-compliant representation of expected responses 237
10.2.2.3 Pedagogical and linguistic notions underlying the RSL 237
10.2.3 Customisable modelling of correct and incorrect responses . . 238
10.2.3.1 Modelling exact matching responses 239
10.2.3.2 Modelling partial matching responses 241
10.2.3.3 Transformation operations 241
10.3 A methodology for teachers to author ICALL materials 245
10.3.1 FL learning activities that suit NLP 245
10.3.2 ReSS: Response Specification Scheme 246
10.3.2.1 RIF-based characterisation of an activity 246
10.3.2.2 Applying the ReSS to a set of expected responses . . 247
10.4 Generating activity-specific NLP resources 251
10.4.1 Exact matching responses 251
10.4.2 Pre-envisaging deviating responses 252
10.4.2.1 Variation derived from omission 252
10.4.2.2 Variation derived from addition 253
10.4.2.3 Variation derived from substitution 257
10.4.2.4 Variation derived from reordering 259
10.4.2.5 Variation derived from blends 260
10.5 Chapter summary 262

11 Integrating ICALL in secondary education environments 263
11.1 Experiment setup . . . 264
11.1.1 Characterisation of the instruction setting . . . 264
11.1.1.1 Expected user actions and roles . . . 265
11.1.2 Participants . . . 266
11.1.2.1 Teacher profiles . . . 266
11.1.2.2 Learner profiles . . . 268
11.1.3 An authoring tool for ICALL activities . . . 269
11.1.3.1 Graphical User Interface . . . 269
11.1.3.2 Automatic generation of NLP resources . . . 276
11.2 Procedure . . . 280
11.2.1 Teacher training . . . 280
11.2.1.1 Introduction to the experiment . . . 280
11.2.1.2 Pedagogical background and activity design . . . 281
11.2.1.3 Automatic feedback for assessment purposes . . . 281
11.2.1.4 Managing AutoTutor . . . 282
11.2.2 Material creation process . . . 282
11.2.3 Application of materials in class . . . 282
11.3 Results . . . 282
11.3.1 Authored materials . . . 282
11.3.1.1 Integration in course programme . . . 283
11.3.1.2 Input data . . . 285
11.3.1.3 TAF characterisation . . . 286
11.3.1.4 RIF characterisation . . . 287
11.3.2 ReSS-based specifications by teachers . . . 289
11.3.2.1 Specifications by T1 . . . 289
11.3.2.2 Specifications by T2 and T3 . . . 291
11.3.2.3 Overview of the complexity of response specifications 295
11.3.3 Use of materials by learners . . . 295
11.3.3.1 Use of materials by T1 . . . 295
11.3.3.2 Use of materials by T2/3 . . . 295
11.3.4 Quality and usefulness of the feedback . . . 296
11.3.4.1 Criteria for the evaluation of feedback . . . 297
11.3.4.2 Feedback to step one in the correction process . . . 299
11.3.4.3 Feedback to step two in the correction process . . . 299
11.3.4.4 Error analysis of the system's performance . . . 300
11.3.4.5 Learner uptake . . . 302
11.4 Discussion . . . 307
11.4.1 Teacher perspective . . . 307
11.4.1.1 AutoTutor's feedback compared to teacher feedback . 308
11.4.2 Learner perspective . . . 311
11.4.3 Research perspective . . . 312
11.4.3.1 Material creation: the process and the product . . . 312
11.4.3.2 Materials used in class . . . 316
11.4.3.3 The limits of AutoTutor's NLP-based feedback . . . 317

11.5 Chapter summary . . . 319

V Conclusions **323**

12 Conclusions and outlook **327**
12.1 Contributions . . . 327
12.1.1 Connecting TBLT and NLP principles . . . 327
12.1.2 NLP as an enabling technology for teachers . . . 329
12.1.3 General contributions . . . 331
12.2 Future work . . . 331
12.2.1 Thesis-related short term research . . . 331
12.2.2 Longer term research in ICALL . . . 332

Appendixes **334**

A ALLES learning units: final tasks and task sequencing **337**
A.1 Career Management and Human Resources . . . 338

B On finite state machines **341**

C ALLES materials as presented to learners **343**
C.1 Screen captures of *Stanley Broadband customer satisfaction questionnaire* . . . 343
C.2 Screenschots of *Describe the structure of your company to a colleague of yours* . . . 346
C.3 Screenshots of *Registering for a course* . . . 349
C.3.1 Input data included in the activity . . . 353
C.3.1.1 "Email from the Human Resources Department' . . . 353
C.3.1.2 "Message from your manager" . . . 353
C.4 Screenshots of *Expresa tu satisfacción o insatisfacción con el producto* Smint . . . 354

D Lexical measures for the assessment of specific vocabulary **357**

E Detailed NLP specifications for activities of Type I, III and IV **359**
E.1 NLP specifications for Task *Customer Satisfaction and International Communication* . . . 359
E.1.1 Specified correct well-formed responses for Item 1 . . . 359
E.1.2 Variations on specified responses for Item 1 . . . 359
E.1.3 Specified correct well-formed responses for Item 2 . . . 359
E.1.4 Variations on specified responses for Item 2 . . . 360
E.1.5 Specified correct well-formed responses for Item 3 . . . 360
E.1.6 Variations on specified responses for Item 3 . . . 360
E.1.7 Specified correct well-formed responses for Item 4 . . . 360
E.1.8 Variations on specified responses for Item 4 . . . 361

E.1.9 Specified correct well-formed responses for Item 5 361
E.1.10 Variations on specified responses for Item 5 361
E.2 NLP specifications for Task *Registering for a course* 362
E.2.1 Specified correct well-formed versions of component "Greeting" 362
E.2.2 Variations on specified versions of component "Greeting" . . . 362
E.2.3 Specified correct well-formed versions of component "IntroYourself" 362
E.2.4 Variations on specified versions of component "IntroYourself" 362
E.2.5 Specified correct well-formed versions of component "YourDept" 362
E.2.6 Variations on specified versions of component "YourDept" . . 362
E.2.7 Specified correct well-formed versions of component "Course" 363
E.2.8 Variations on specified versions of component "Course" 363
E.2.9 Specified correct well-formed versions of component "Schedule" 363
E.2.10 Variations on specified versions of component "Schedule" . . . 364
E.2.11 Specified correct well-formed versions of component "AuthorisedBy" 364
E.2.12 Variations on specified versions of component "AuthorisedBy" 364
E.2.13 Specified correct well-formed versions of component "UsefulFuture" 364
E.2.14 Variations on specified versions of component "UsefulFuture" . 364
E.2.15 Specified correct well-formed versions of component "FutureInterest" 365
E.2.16 Variations on specified versions of component "FutureInterest" 365
E.2.17 Specified correct well-formed versions of component "ComplClose" 365
E.2.18 Variations on specified versions of component "ComplClose" . 365
E.2.19 Specified correct well-formed versions of component "Signature" 365
E.2.20 Variations on specified versions of component "Signature" . . 365
E.3 NLP specifications for Task *Expresa tu satisfacción o insatisfacción con el producto* Smint 366
E.3.1 Specified correct well-formed versions of component "Saludo" . 366
E.3.2 Variations on specified versions of component "Saludo" 366
E.3.3 Specified correct well-formed versions of component "RazonCarta" 366
E.3.4 Variations on specified versions of component "RazonCarta" . 366
E.3.5 Specified correct well-formed versions of component "Opinion" 366
E.3.6 Variations on specified versions of component "Opinion" . . . 366
E.3.7 Specified correct well-formed versions of component "MasInfo" 367
E.3.8 Variations on specified versions of component "MasInfo" . . . 367
E.3.9 Specified correct well-formed versions of component "Despedida" 367
E.3.10 Variations on specified versions of component "Despedida" . . 367
E.3.11 Specified correct well-formed versions of component "Firma" . 367
E.3.12 Variations on specified versions of component "Firma" 367

F TAF and RIF analysis of the E.T. activity **369**
F.1 TAF analysis . . . 369
F.2 RIF analysis . . . 370

G Teacher traning material in ICE3 **371**
G.1 Table for the characterisation of CALL/ICALL activities . . . 371

H Formal analysis of the ICALL activities authored by teachers **373**
H.1 Teacher 1's work plan . . . 373
H.2 Chemical reactions – Describing reactants and products . . . 377
H.2.1 TAF analysis . . . 379
H.2.2 RIF analysis . . . 380
H.2.2.1 Detailed RIF analysis for Item 1 . . . 381
H.3 Chemical reactions – Calculating theoretical yields . . . 382
H.3.1 TAF analysis . . . 383
H.3.2 RIF analysis . . . 384
H.3.2.1 Detailed RIF analysis for Item 1 . . . 385
H.4 Analysis of graphs (II) . . . 386
H.4.1 TAF analysis . . . 387
H.4.2 RIF analysis . . . 388
H.4.2.1 Detailed RIF analysis for Item 1 . . . 389
H.5 Daily routines II . . . 390
H.5.1 TAF analysis . . . 391
H.5.2 RIF analysis . . . 392
H.5.2.1 Detailed RIF analysis for Item 1 . . . 392
H.6 The good and the bad student . . . 394
H.6.1 TAF analysis . . . 395
H.6.2 RIF analysis . . . 395
H.6.2.1 Detailed RIF analysis for Item 1 . . . 397
H.7 Detailed ReSS specifications . . . 398
H.8 Complexity of ReSS specifications . . . 428

Bibliography **447**

List of Figures

2.1 General architecture of an Intelligent Language Tutoring System – simplification of the one proposed in Amaral (2007: p. 85). 20
2.2 The viable processing ground (Bailey and Meurers, 2008: p. 108). . . 28

3.1 Full syntactic parse of a sentence in parenthetic and tree representation. 41
3.2 Partial syntactic parse of a sentence in parenthetic representation. . . 42

4.1 Linguistic content specified for a task according to the approach proposed by Estaire and Zanón (1994: p. 63). 59

5.1 Differences in teacher/learner and virtual tutor/learner interaction during the learning. 76
5.2 The ICALL task life cycle. 77
5.3 Processes and interrelationships within the ICALL activity focused on for the definition and exemplification of the methodology proposed. . 81

6.1 ALLES topics according to interest area and learner CEF level. . . . 86
6.2 A modular and domain-adaptive NLP architecture with for the processing of learner responses. 93
6.3 Analysis of the German sentence *Der Weg ist frei* by the MPRO module in the MPRO-KURD linguistic annotation solution. 97
6.4 Analysis of the German sentence *Der Weg ist frei* by the KURD-based Morphological Disambiguator. 98
6.5 KURD rule to disambiguate the readings of the words *Der Weg* to their nominative masculine singular readings in the sentence *Der Weg ist frei*. 98
6.6 Results of the tokenisation and morphological analysis process for the Catalan sentence *La casa és verda*. 100
6.7 Disambiguation rule that applies to the word *La* to remove the pronoun reading in the analysis of the sentence *La casa és verda*. 101
6.8 Disambiguation rule applying to the word *casa* to select the noun reading in the analysis of the sentence *La casa és verda*. 101
6.9 CG rules for the analysis of the complimentary close in a formal letter in Catalan. 102
6.10 KURD rules to process a part of a possible response to one of the ICALL activities later on presented and worked out in Chapters 7 and 8. 103

6.11 Linguistic analysis for the sentence (5) including the response elements detected by the Information Extraction Module. 104

7.1 Four task types in the viable processing ground. 151

8.1 NLP specification procedure. 155
8.2 Specification procedure for the feedback generation logic. 155
8.3 System-learner interaction for the evaluation of responses with the Automatic Assessment Module. 173
8.4 Two-step feedback presentation flux in ALLES. 174
8.5 A domain-adaptive NLP-based feedback generation architecture for formative and summative assessment. Dotted lines indicate domain-specific resources. 175
8.6 Recognition paths to analyse linguistic structures relevant for the assessment of thematic and linguistic contents. 177
8.7 Global response evaluation recognition path for the response to Item 1 in the customer-satisfaction-questionnaire activity. 178
8.8 KURD rules to process a part of a possible response to one of the ICALL activities later on presented and worked out in Chapters 7 and 8. 178
8.9 Recognition path to analyse a response including wrong word form errors (I). 180
8.10 Recognition path to analyse a response including wrong word form errors (II). 181
8.11 Recognition path to analyse a response including wrong lexical choice errors. 183
8.12 Graphs representing the recognition paths for modelling response components with missing or unexpected information. 185
8.13 Graphs representing the recognition paths for global response evaluation of responses with missing or unexpected response components. . 186
8.14 Recognition path of the response components of a language learning activity with an extended production response. 187
8.15 Partial recognition path of the element "Course and availability" for the response to the Final Task of the learning unit *Education and Training*. 188
8.16 Flexible recognition paths using a "bag-of-words" approach to response component analysis. 189
8.17 CG-based rule in the Global Response Checker to be applied to the Catalan version of the activity described in section 7.2.2.3. 189
8.18 Language content recognition paths for loosely restricted responses. . 190
8.19 Recognition path of the response components of a task with a loosely restricted response – Activity 5 in Subtask 2 of *Antención al cliente* in ALLES. 191
8.20 Indicators obtained from the learner response for summative assessment. 193
8.21 Example of summative feedback obtained with the current version of the ALLES system. 194

9.1 Recognition paths required for a finer grained feedback to preposition errors. 225

10.1 Context of use of an ICALL activity authoring and management tool including NLP as an enabling technology. 233
10.2 An customisable NLP architecture for the processing of responses to ICALL activities designed by FLTL practitioners. 235
10.3 Abstract list of Response Components and the corresponding RC Sequences. 238
10.4 Customised FSA recognition paths derived from the RSL specifications in Figure 10.3. 240
10.5 Expansion of RSL-based response patterns using omission transformation operations. 242
10.6 Expansion of RSL-based response patterns using substitution transformation operations. 243
10.7 Expansion of RSL-based response patterns using addition transformation operations. 243
10.8 Expansion of RSL-based response patterns using reordering as a transformation operation. 244
10.9 Expansion of RSL-based response patterns using blending as a transformation operation. 245
10.10 Response Components of the response to the E.T. comprehension activity. 247
10.11 Identification and classification of the Response Components. 248
10.12 Graphical representation of the Response Components and the correct RC Sequences for the E.T. comprehension activity. 250
10.13 FSA recognition paths generated for the strings of Variants B1 and B2 in the RC E.T. REPEATING WORDS. 251
10.14 FSA recognition paths generated for the checking of exact matching RC sequences for the E.T. activity. 252
10.15 Expansion of RSL-based response patterns using omission as a transformation operation. 254
10.16 Recognition paths expanding one of the RC sequences of the E.T. comprehension activity. 255
10.17 Paths generated by expanding B1 variant strings by means of addition operations. 255
10.18 Paths generated by expanding B1 variant strings by means of POS-filtered addition operations. 256
10.19 Paths generated by expanding variant strings and RC sequences by means addition of unexpected linguistic items. 256
10.20 Paths generated by expanding B1 variant strings by means of addition operations. 257
10.21 Paths generated by expanding B1 variant strings by means of substitution operations. 258
10.22 Paths generated by expanding C2 variant strings by means of reordering operations. 259

10.23 Recognition paths expanding one of the RC Sequences of the response to the E.T. comprehension activity. 260
10.24 Paths generated by expanding C2 variant strings by means of blending as a transformation operation. 261
10.25 Recognition paths expanding one of the RC sequences from Figure 10.14 by allowing blending structures. 261

11.1 ATACK's GUI: settings and global activity actions 270
11.2 ATACK's GUI: activity authoring area 271
11.3 ATACK's GUI: question tab area to specify information for response components. 272
11.4 ATACK's GUI: specification of the information for response component E.T. REPEATS WORDS. 273
11.5 ATACK's GUI: ordering of the response components to build the sequences yielding correct responses. 274
11.6 ATACK's GUI: *Customised feedback* tab, an area to define learner-oriented feedback messages. 275
11.7 ATACK's GUI: *Sample answers* tab, an area to insert sample answers to be shown as part of the feedback. 276
11.8 From the specifications provided by content designers in ATACK to the NLP resources needed by ATAP to evaluate learner responses. . . 277
11.9 Software for the generation of expanded form-based recognition paths in the Content Analyser. 278
11.10 Fragment of the workplan designed by T1 to include ICALL and CALL materials in the course sessions on chemical reactions. 283
11.11 Fragment of the workplan designed by T1 to include ICALL and CALL materials in the course sessions on chemical reactions. 284
11.12 Overview of the materials created by T2 and T3 for the ESL courses. 285
11.13 ReSS specifications for Item 1 in A1-T1. 291
11.14 ReSS specifications for Item 1 in A1-T1. 294
11.15 Feedback messages generated for a learner response that include some misleading information. 302
11.16 Learner uptake for valid feedback messages to spelling and grammar errors. 305
11.17 Learner uptake for valid feedback messages to activity specific language and content errors. 306
11.18 Learner responses corrected manually by T2. 309
11.19 Agreement between system's feedback and teacher comments. 310
11.20 Satisfaction of learners with AutoTutor activity feedback. 312
11.21 Limitations of current feedback strategy Feedback exemplified on a real learner response. 319

C.1 Screen capture of the overview of the task "Stanley Broadband customer satisfaction questionnaire". 344
C.2 Screen capture of the details of the task "Stanley Broadband customer satisfaction questionnaire" (I). 345

C.3 Screen capture of the details of the task "Stanley Broadband customer satisfaction questionnaire" (II). 345
C.4 Screen capture of the overview of the task "Describe the structure of your company to a colleague of yours", Activity no. 5. 346
C.5 Screen capture of the details of the task "Describe the structure of your company to a colleague of yours", Activity no. 5. 347
C.6 Screen capture of the overview of the task "Describe the structure of your company to a colleague of yours", Activity no. 6. 348
C.7 Screen capture of the details of the task "Describe the structure of your company to a colleague of yours", Activity no. 6. 348
C.8 Screen capture of the overview of the task "Registering for a course", Activity no. 1. 349
C.9 Screen capture of the details of the task "Registering for a course", Activity no. 1. 350
C.10 Screen capture of the overview of the task "Registering for a course", Activity no. 2. 351
C.11 Screen capture of the details of the task "Registering for a course", Activity no. 2. 352
C.12 Screen capture of the email given as input data to the learner in Task "Registering for a course". 353
C.13 Screen capture of the overview of the task "Expresa tu satisfacción o insatisfacción con el producto *Smint*". 354
C.14 Screen capture of the details of the task "Expresa tu satisfacción o insatisfacción con el producto *Smint*" (I). 355

H.1 Screen capture of the activity "Chemical reactions – Reactants and Products" created by T1. 378
H.2 Screen capture of the activity "Chemical reactions – Changing rates" created by T1. 382
H.3 Screen capture of the activity "Analysis of graphs (II)" created by T1. 386
H.4 Activity "Daily routines II" created by T2 and T3. 390
H.5 Screen capture of the activity "The good and the bad student" created by T2 and T3. 394
H.6 ReSS specification RC and RCS for question 1 in activity AnalysisOfGraphs-v1 by T1. 398
H.7 ReSS specification RC and RCS for question 2 in activity AnalysisOfGraphs-v1 by T1. 399
H.8 ReSS specification RC and RCS for question 3 in activity AnalysisOfGraphs-v1 by T1. 400
H.9 ReSS specification RC and RCS for question 4 in activity AnalysisOfGraphs-v1 by T1. 401
H.10 ReSS specification RC and RCS for question 5 in activity AnalysisOfGraphs-v1 by T1. 402
H.11 ReSS specification RC and RCS for question 6 in activity AnalysisOfGraphs-v1 by T1. 403

H.12 ReSS specification RC and RCS for question 7 in activity AnalysisOfGraphs-v1 by T1. 404
H.13 ReSS specification RC and RCS for question 8 in activity AnalysisOfGraphs-v1 by T1. 405
H.14 ReSS specification RC and RCS for question 1 in activity Changingtherates1 by T1. 406
H.15 ReSS specification RC and RCS for question 2 in activity Changingtherates1 by T1. 407
H.16 ReSS specification RC and RCS for question 3 in activity Changingtherates1 by T1. 408
H.17 ReSS specification RC and RCS for question 4 in activity Changingtherates1 by T1. 409
H.18 ReSS specification RC and RCS for question 5 in activity Changingtherates1 by T1. 410
H.19 ReSS specification RC and RCS for question 6 in activity Changingtherates1 by T1. 411
H.20 ReSS specification RC and RCS for question 7 in activity Changingtherates1 by T1. 412
H.21 ReSS specification RC and RCS for question 8 in activity Changingtherates1 by T1. 413
H.22 ReSS specification RC and RCS for question 1 in activity ChemicalReactions-v1 by T1. 414
H.23 ReSS specification RC and RCS for question 2 in activity ChemicalReactions-v1 by T1. 415
H.24 ReSS specification RC and RCS for question 3 in activity ChemicalReactions-v1 by T1. 416
H.25 ReSS specification RC and RCS for question 4 in activity ChemicalReactions-v1 by T1. 417
H.26 ReSS specification RC and RCS for question 5 in activity ChemicalReactions-v1 by T1. 418
H.27 ReSS specification RC and RCS for question 6 in activity ChemicalReactions-v1 by T1. 419
H.28 ReSS specification RC and RCS for question 7 in activity ChemicalReactions-v1 by T1. 420
H.29 ReSS specification RC and RCS for question 1 in activity PerfectStudent-v1 by T2/T3. 421
H.30 ReSS specification RC and RCS for question 2 in activity PerfectStudent-v1 by T2/T3. 422
H.31 ReSS specification RC and RCS for question 3 in activity PerfectStudent-v1 by T2/T3. 423
H.32 ReSS specification RC and RCS for question 4 in activity PerfectStudent-v1 by T2/T3. 424
H.33 ReSS specification RC and RCS for question 1 in activity Routines1 by T2/T3. 425

H.34 ReSS specification RC and RCS for question 4 in activity Routines1 by T2/T3. 426
H.35 ReSS specification RC and RCS for question 7 in activity Routines1 by T2/T3. 427

List of Tables

6.1 Partners of the ALLES consortium and their respective expertise. . . 84
6.2 Sentence annotated with the initial levels of analysis of the proposed architecture. 95

7.1 Subtask 0 of the learning unit *Education and Training*. 107
7.2 Subtask 1 of the learning unit *Education and Training*. 110
7.3 Subtask 2 of the learning unit *Education and Training*. 112
7.4 Subtask 3 of the learning unit *Education and Training*. 114
7.5 Final Task of the learning unit *Education and Training*. 115
7.6 Partial RIF characterisation of Activity 4 in Subtask 3 in *Customer Satisfaction and International Communication*. 121
7.7 RIF analysis of Item 1 of Activity 4 in Subtask 3 of the unit *Customer Service and International Communication* in ALLES. 122
7.8 Application of the RIF to Activities 5 and 6 in Subtask 1 in *Company Organisation* in ALLES. 126
7.9 Application of the RIF analysis to Activities 5 and 6 in Subtask 1 in the unit *Company Organization* in ALLES. 129
7.10 Application of the RIF to Activities 1 and 2 in the Final Task in *Education and Training* in ALLES. 134
7.11 Thematic and linguistic content according to the RIF for Activities 1 and 2 in the Final Task in *Education and Training*. 136
7.12 Indicators for the assessment of communicative contents: thematic content (TC) and linguistic content (LC) at the level of text genre. . 139
7.13 Indicators for the assessment of lexical contents: use of specific vocabulary (SV) and the number of words (NW) as a fluency measure. . . . 140
7.14 Indicators for the assessment of sentence structure and accuracy: number of sentences (NS), number of discourse markers (NDM), and number of grammar and usage errors (NGE) in the response. 142
7.15 Table co-relating the number of paragraphs (NP) and the number of spelling errors (NSE) in the response. 143
7.16 Application of the RIF to Activity 5 in Subtask 2 in *Atención al cliente*.146
7.17 Thematic and linguistic content according to the RIF for Activity 5 in Subtask 2 in *Atención al cliente*. 148

8.1 SALA applied to the thematic content part of the RIF for Item 1 of Activity 4 in Subtask 3 of the Learning Unit *Customer Service and International Communication* in ALLES. 157

8.2 SALA applied to the linguistic content part of the RIF for Item 1 of Activity 4 in Subtask 3 of the Learning Unit *Customer Service and International Communication* in ALLES. 159
8.3 SFGL applied to the thematic content part of the RIF for Item 1 of Activity 4 in Subtask 3 of the Learning Unit *Customer Service and International Communication* in ALLES. 161
8.4 SGFL applied to the linguistic content part of the RIF for Item 1 of Activity 4 in Subtask 3 of the Learning Unit *Customer Service and International Communication* in ALLES. 164
8.5 SALA applied to the analysis requirements for the thematic contents in tasks requiring summative assessment. 166
8.6 SALA applied to the analysis requirements for the linguistic contents in tasks requiring summative assessment. 168
8.7 SALA applied to the analysis requirements for the application of CAF measures in tasks requiring summative assessment. 169
8.8 Application of the AASF to the Activities 1 and 2 in the Final Task in *Education and Training* in ALLES. 172
8.9 Abstract representation of the linguistic analysis obtained with the general module of the architecture. 175
8.10 Comparison of the analysis of Example (11a) with the one of the correct response *How satisfying are you with StanleyBroadband?* . . . 179
8.11 Comparison of the analysis of Example (11b) with the one of the correct response *How satisfying are you with StanleyBroadband?* . . . 181
8.12 Comparison of the correct response with the analysis of *How satisfied are you* against *Stanley Broadband?* 182

9.1 Classes of variation obtained by crossing the criteria of correctness and well-formedness. 197
9.2 Manually annotated well-formed variations per response component. . 200
9.3 Distribution of correct and incorrect responses to the items in Activity 4 of Subtask 3 in *Customer Satisfaction and International Communication*. 202
9.4 Well-formed and ill-formed structures in correct and incorrect responses to the items in Activity 4 of Subtask 3 in *Customer Satisfaction and International Communication* in ALLES. 203
9.5 Responses using linguistic structures that match or diverge from design specifications for Activities 1 and 2 of the Final Task in *Education and Training*. 207
9.6 Distribution of correct and incorrect responses to Activities 1 and 2 of the Final Task in *Education and Training*. 208
9.7 Ill-formed structures in correct and incorrect responses to the items in Activity 4 of Subtask 3 in *Customer Satisfaction and International Communication*. 210
9.8 Responses using linguistic structures that match and diverge from design-based specifications for Activity 5 of Subtask 2 in *Atención al cliente* in ALLES. 214

9.9 Distribution of correct and incorrect responses to Activity 5 of Subtask 2 in *Atención al cliente* in ALLES. 215
9.10 Ill-formed variation annotations in correct and incorrect responses to the items in Activity 5 of Subtask 2 in *Atención al cliente* in ALLES. 216
9.11 Well-formed variations for the responses to the items in Activity 5 of Subtask 2 in *Atención al cliente* in ALLES. 219
9.12 Response length average in words per activity type. 221
9.13 Percentage of response element annotations indicating unexpected, missing or expected content. 222

10.1 Formal definition of the Response Specification Language. 237
10.2 Pedagogically and linguistically relevant concepts of the Response Specification Language. 237

11.1 Heuristics for the selection of linguistic features in the reference texts for the expansion of the response components 279
11.2 Number of RC, RC Variants and Strings, RCS and total number of generated sentences for activity items in A1-T1. 292
11.3 Number of RC, RC Variants and Strings, RCS and total number of generated sentences for activity items in A1-T2/3. 293
11.4 Total number of Response Components, Variants and Strings per question in the ICALL activities authored by teachers. 295
11.5 Response attempts by learners in 3A, 3B and 3C working with materials generated by T1. 296
11.6 Response attempts by learners in 1A, 1B, 2A and 2B working with materials generated by T2/3. 297
11.7 Validation strategy for the evaluation of feedback quality in terms of spelling and grammar. 298
11.8 Validation strategy followed for the evaluation of system performance in terms of task-specific content and language. 298
11.9 Goodness of ICALL feedback in spell and grammar checking. 299
11.10 Goodness of ICALL feedback in activity specific language and content checking. 300
11.11 Frequency and nature of the messages that yielded wrong explanations in responses that included real errors or deviations. 301
11.12 Analysing learner take up on the basis of changes in resubmissions for the correction of spelling and grammar errors. 303
11.13 Analysing learner take up on the basis of changes in resubmissions for the correction of activity specific language and content. 304
11.14 Excerpt of the system-teacher comparison register. 309

F.1 RIF characterisation of the activity "E.T. – The Extraterrestrial". . . 370

G.1 Table for teachers to characterise their learning activities. 371

H.1 RIF characterisation of the activity "Chemical reactions – Reactants and products" by Teacher 1. 380

H.2 Detailed RIF analysis for Item 1 in the Activity "Chemical reactions – Reactants and products". 381
H.3 RIF characterisation of the activity "Chemical reactions – Changing rates" by Teacher 1. 384
H.4 Detailed RIF analysis for Item 1 in the Activity "Chemical reactions – Changing rates". 385
H.5 RIF characterisation of the activity "Chemical reactions – Changing rates" by Teacher 1. 388
H.6 Detailed RIF analysis for Item 1 in the Activity "Analysis of graphs (II)". 389
H.7 RIF characterisation of the activity "Daily routines II" by Teachers 2 and 3. 392
H.8 Detailed RIF analysis for Item 1 in the Activity "Daily routines II". . 393
H.9 RIF characterisation of the activity "The good and the bad student" by Teachers 2 and 3. 396
H.10 Detailed RIF analysis for Item 1 in the Activity "The good and the bad student". 397
H.11 Total number of Response Components, Variants and Strings per question in the ICALL activities authored by teachers. 428

Part I

Introduction

Scientists and organizations should consider the benefits and costs of collaboration before deciding to collaborate. Collaboration only for the sake of collaboration does not seem warranted given the number of factors that should be taken into account before and during a collaboration. Furthermore, as the number and diversity of participants and the complexity and uncertainty of the scientific work increase, so does the complexity of the factors. The negative consequences from not addressing the factors may also increase. There is a real need to consider these factors and the effort and other costs required to manage them before beginning a collaboration. However, when collaboration can provide new possibilities, it is well worth the effort. New possibilities offered by collaboration can be many and diverse, including new ways of conducting science and new knowledge to the benefit of many.

Scientific collaboration
A synthesis of challenges and strategies
Diane H. Sonnenwald (2007: p. 671–672)

Chapter 1

Motivation and research goals

This thesis studies the application of computer-based language processing to foreign language teaching and learning. In particular, we investigate the use of Computer-Assisted Language Learning (CALL) materials including Natural Language Processing (NLP) techniques for the provision of automated feedback to learners in Foreign Language Teaching and Learning (FLTL). This is a field traditionally known as Intelligent CALL (ICALL). As we will see through this introductory part, ICALL research is not free of controversy. NLP researchers and FLTL and CALL researchers have defended opposite claims regarding its usefulness and its feasibility, especially regarding its usefulness in real-life instruction settings and its feasibility in terms of the NLP functionalities required. This thesis attempts to connect these two worlds in theory and in practice.

This thesis has two main research goals. On one hand, this thesis aims to show and justify that ICALL activity design and implementation is a cross-disciplinary process that demands a tight relationship between pedagogical needs and computational capabilities. To elucidate this relationship we explore three main areas: (i) the characteristics of Foreign Language (FL) learning activities, i.e. pedagogical goals; (ii) the complexity of the linguistic elements in the elicited responses, i.e. learner language; and (iii) the feasibility of applying NLP techniques for the automatic evaluation of learner responses, i.e. computational capabilities.

On the other hand, this thesis studies the application of NLP-enhanced technologies for FLTL teachers to be able to autonomously design and use ICALL activities in instruction settings. To achieve this goal we develop new methodological and technical concepts for the authoring of ICALL activities, which we test in an experiment with secondary school teachers. The experiment includes the design, creation and use of ICALL materials by the teachers in the respective instruction settings. Additionally, a quantitative and qualitative evaluation is performed from the point of view of the teacher and the learner, accompanied by a discussion from the perspective of the research.

In this introductory part, including Chapters 1 and 2, we define the research context in which this work is to be framed, and the questions that motivate it. At the end of these chapters, we will revisit and further specify the thesis goals.

1.1 Computers in foreign language learning

In his *Principles Of Language Learning and Teaching*, Brown (2007) states that the acquisition of a foreign language is, among other things, a process that results from a variety of individual and social activities, of which language instruction can be a part. In the acquisition of a foreign language "instruction makes a difference in learner's success rates" (Brown, 2007: p. 269), a thesis which is supported by evidence from research (e.g., Buczowska and Weist, 1991; Doughty, 1991, 2003; Ellis, 2005).

As with many other professional activities, education has seen how computers have become an essential instrument for daily practice since the late 1980s (Levy, 1997: p. 15). Teachers and learners have since become more and more aware of the possibilities made available by the use of computers. However, long before the 1980s the FLTL community was experimenting with the use of computers for language learning – it all started back in the 1960s (*ibid.*).

Levy defines Computer-Assisted Language Learning (CALL) as "the search for and study of applications of the computer in language teaching and learning" (1997: p. 1). According to Levy and Stockwell (2006: p. 2), CALL is a means of enhancing foreign language teaching and learning in that it allows learners to manipulate language more effectively, supplies context-sensitive information, and offers greater flexibility. In his historical overview of CALL, Levy (1997: ch. 2) reviews several successful projects in CALL since the 1960s. In line with Pederson (1988) and Dunkel et al. (1991), Levy concludes that CALL can really encourage and increase the motivation and success of language learners (1997: p. 29). He also emphasises that "meaningful (as opposed to manipulative) CALL practice is both possible and preferable" both from the teacher and from the learner perspectives (ibid.).

The most common use of computers in FLTL is to facilitate the delivery of materials in formats other than the printed paper (video, audio) and provide learners and teachers with a channel of communication that allows for distant and interactive activities (from hypertext to virtual realities, through email, chat or video chat functionalities). However, computers can be used as a interaction agent, an agent enhanced with artificial intelligence.

The integration of Artificial Intelligence, and particularly of NLP, in CALL systems is commonly known as Intelligent CALL (ICALL). Heift and Schulze (2007: p. 56) identify more than one hundred projects focusing on the integration of NLP until 2004. As we will see, although the commonalities between CALL and ICALL are more in number than the differences, many of the works in each field all too often paid too little attention to the other field (Schulze, 2010: p. 79). According to Heift and Schulze (2007), during the last 40 years, research and development in ICALL has focused on the design and implementation of robust NLP techniques for the automatic processing of learner language, on the description of learner language – including and often limited to error analysis –, on the generation of feedback coherent with the pedagogical approach and the learner needs, and on the modelling of students with the aim to build intelligent language tutoring systems.

To sum up: the use of computers in FLTL (CALL or ICALL) has a relatively short tradition. However, over 50 years of research in CALL have made it clear that this is a complex and interdisciplinary task. CALL and ICALL research demand expertise in fields such as Linguistics, Psychology, Educational Technology, User Modelling, Computational Linguistics, and, of course, Second Language Acquisition, and Foreign Language Teaching and Training. This is the context in which the first motivation of this thesis shows up, namely the goal to carry out multidisciplinary research integrating theory and practice in both FLTL and NLP.

1.2 NLP as a transferable technology

The automatic analysis of learner language in ICALL is possible thanks to the application of Natural Language Processing techniques, a term often used as a synonym of Computational Linguistics (CL).[1] CL is mainly concerned with the formalisation and validation of the procedures involved in the processing of human language, including computer-based simulations. CL has both a theoretical and a practical side: The former is concerned with the description and representation of formalisms with which human language abilities can be emulated. This theoretical side studies formal grammars, the representation of linguistic information, and the algorithms required to both parse and generate language by means of a computer. The more practical side of CL is concerned with the application of language processing software to solve language-mediated tasks such as speech recognition, machine translation, automatic summarisation, or language checking.

As argued in ten Hacken (2003), the successful application of NLP-based software relies on the ability to find "a good problem to solve". This apparently trivial statement alludes to the need of finding a real problem with concrete users that have concrete goals. In ten Hacken's view, not being able to find the good problem to solve is what explains most of the dissatisfaction and frustration in Machine Translation (MT), the oldest and most popular field of application of CL, until the 1980s. Only after the research community starts to look at the problem of MT as a particular and well-defined communicative human activity, one in which two parties are truly interested in the message coming through, can this frustration be overcome. In his view, this also explains the success of MT systems such as Météo (Isabelle, 1987; Chandioux, 1989), which translated meteorological reports between several languages.

In spite of four decades of research in ICALL, Heift and Schulze (2007: pp. 224–225) suggest there is a reciprocal frustration and disenchantment both among NLP developers and among FLTL and SLA researchers with regard to its use and effectiveness. NLP researchers and developers tend to think that CALL underuses NLP and that its potential is not well seized for foreign language learning and teaching. FLTL and SLA researchers, as well as FL teachers, tend to think that NLP is still too immature to be meaningfully used in modern foreign language teaching.

Even though NLP is far from offering full-fledged human language processing, many technologies in our daily life include NLP-based solutions: word guessers in mobile devices, spam filters, automatic email classification, spell and grammar check-

[1] Natural Language Processing is generally the name used by researchers in the tradition of Computer Science, while Computational Linguistics is typically used by researchers in the tradition of Linguistics. In this thesis we use both terms indistinguishably, except if explicitly noted.

ers in text processors, search engines, natural language interfaces to data bases, phone banking and customer services using speech recognition, reading assistants including speech synthesis, etc. If the application of NLP-enhanced technologies has made it into real-life tasks, it seems reasonable to think that finding a good problem to solve in the context of language teaching and learning should also allow us to develop effective and practical ICALL systems.

Schoelles and Hamburger (1996) suggest that in this respect NLP faces a typical situation of technology transfer, which resembles an elections procedure: many candidates, many technologies, competing to draw the attention from the interested parties for a limited number of positions. Each candidate has to convince the interested parties (in CALL that would be teachers, learners, SLA researchers, educational managers, etc.) that his or her proposal is useful and ideally better than the current state of affairs. In this context, competition exists between NLP-aware and NLP-unaware solutions.

Our belief is that technology transfer requires a state of mutual understanding between technology creators and technology users. In our view, this state of mutual understanding can only be reached by transferring two important sources of knowledge. The first source of knowledge is the experience of the experts and the researchers that work daily in the setting in which the technology will be eventually integrated. Before technology becomes a truly enabling technology – as cars can be driven without any knowledge of mechanics, physics, or electronics –, technology creators need to learn as much as possible about the actual conditions and needs of the professional environment in question: how the participants in this environment think and work, what their main goals and their main fears are, which things are important to them and which things are less important or irrelevant, which tools they work with and which tools they plan to work with, and, of course, what costs they are committed to assume – personal, collective, organisational or economical. In our research this knowledge mainly corresponds to research and practice in FLTL.

The second important source of knowledge to be transferred is the capabilities and limitations of the technology. This can be approached in many different ways, depending on factors such as the degree of development of the technology: Is it already there or is it still in the design stage? How much does the technology change the conceptual aspects of the activities performed in the learning environment? Is it an enabling technology that allows for a creative unexpected use? Does the user have a prosumer role, as opposed to a consumer role? Ultimately, the end user has to be able to imagine, experience and evaluate what the technology in question actually enables him or her to do, and how the use of this technology could evolve. In our research this knowledge basically corresponds to research and practice in NLP.

This thesis is motivated by the need to explore the means that would allow us to transfer information between the NLP and the FLTL worlds, its researchers and practitioners. We pursue this goal taking into account methodological and technological views from both worlds.

1.3 The interrelationship between NLP and FLTL

NLP and FLTL have in common that they both work with language. The former works on the automatic processing of it, and the latter works on how language is taught and learnt. Thus, language and linguistics become the natural meeting point for NLP and FLTL.

Yet, Heift and Schulze (2007: p. 221) claim that the complaints about the lack of NLP in CALL-based FLTL and SLA are as frequent as the complaints about the immaturity of NLP – an immaturity that prevents ICALL to be useful in modern FLTL. Researchers in the field provide several reasons that account for this mutual dissatisfaction, but they all agree that there is a profound need to design and develop pedagogically principled ICALL systems (Heift and Schulze, 2007: p. 226, Amaral, 2007: p. 53–63, Amaral and Meurers, 2011: pp. 6 and 12). To overcome this, they advise, research in ICALL should work on the inclusion of NLP technology in settings following communicative approaches to FLTL – particularly task-based approaches.

Communicative approaches to language teaching stress the importance of teaching a language through communicative practice and for communicative purposes (cf., e.g., Richards and Rodgers, 2001, Ellis, 2003: ch. 1, Littlewood, 2004: pp. 320–321). As pointed out by Bailey and Meurers (2009: p. 2), in consequence ICALL systems "should be able to offer a range of contextualized, meaningful language learning activities". However, to do so ICALL systems have to handle both form and meaning, that is, "to be able to recognize multiple realisations of the same meaning, possibly in the presence of form errors" (ibid.).

If, as we suggested, the automatic understanding of unrestricted human language on the basis of NLP techniques is unfeasible, then, as Bailey and Meurers (2009: p. 2) put it, one of the challenges of current ICALL is to better determine the extent to which FLTL needs can be reliably fulfilled by state-of-the-art NLP technology. For such a goal to be achieved, Bailey and Meurers argue for the need to elucidate the relationship between the pedagogical and linguistic characteristics of language learning activities and the degree of variation in learner responses, since this determines the language to be automatically evaluated with NLP technologies.

Thus, this thesis is also motivated by the need to study the methodology that will enable us to characterise FL learning activities and the learner responses to be elicited, and how this characterisation can be used to inform the integration of NLP-based assessment functionalities. In doing so, we commit to the use of modern approaches to foreign language teaching, as well as to the use of state-of-the-art well-known NLP techniques.

1.4 ICALL in real-world instruction settings

Despite the number of research projects carried out over the past years, ICALL still has little presence in real-world instruction settings. Amaral and Meurers (2011: pp. 5–6) state that apart from *TAGARELA*, their system for beginner learners of Portuguese, "there are only two [ICALL] systems that use NLP technology and are fully integrated into real-life foreign language programs in universities". These two systems that they refer to are *Robo-Sensei* (Nagata, 2002), an ICALL system for beginner to intermediate learners of Japanese, and *E-Tutor* (Heift, 1998, 2003), an ICALL system for beginners of German.

As mentioned above, until 2004 more than 100 ICALL projects were identified by Heift and Schulze (2007: pp. 55–56), and most of them had as one of their objectives an evaluation phase with learners. However, strikingly, most of those projects did not get to the point where the developed systems, or even prototypes, could be used with learners and teachers in real-world instruction settings, sometimes not even on controlled experimental settings. The immaturity of the tools has often been an argument, probably a deserved one, for the lack of use of ICALL materials in language learning (Gamper and Knapp, 2002: p. 332).

ICALL researchers provide two further reasons to account for the limited presence of ICALL in foreign language programmes in real-world instruction settings. The first reason is the lack of interdisciplinary ICALL research including SLA and FLTL expertise (Heift and Schulze, 2007: p. 82, Amaral and Meurers, 2011: p. 6–7). Though there is interdisciplinary research in ICALL, many projects were carried out either without a team of language pedagogues or in teams where the interaction between pedagogues and NLP experts did not result into a successful combination of experiences – that is, knowledge transfer did not happen.

The second reason is the absence of teachers, and often learners, in the design, development and evaluation stages of ICALL systems (Heift and Schulze, 2007: p. 226–227, Amaral and Meurers, 2011: p. 6). It is not accidental that the three systems that have actually been introduced in foreign language programmes have been developed by interdisciplinary teams in which FLTL and SLA had as much a weight as NLP had. It is not accidental that these ICALL systems were actually used by teachers who had either created or had an influence on the design of the materials – with their course programmes in mind (Levy, 1997: p. 19). It is not accidental that after several terms (years) of use these systems are now offered to learners and teachers in other course programmes (Nagata, 1997a, 2004; Heift, 1998, 2005; Amaral, 2007). In the words of Heift and Schulze, these systems have managed to turn their work into one that "combines *research* and *development*", theory and practice (2007: p. 9, original italics).

This thesis is thus further motivated by the need to include users – teachers and learners – in ICALL projects (Heift and Schulze, 2007: pp. 222 and 226, Amaral and Meurers, 2011: p. 6), and we do so on the basis of real-life experience working with teachers and students as part of three research projects in which we were involved in the past decade. We investigate the methodology to facilitate the inclusion of ICALL authoring and management tools in real-world instruction settings, with the aim to do it in a usable and effective manner from the perspective of the teacher and the learner.

1.5 Research goals

At this point we want to further detail the two main research goals of the thesis. The first goal of this thesis is to propose and validate a methodology that helps both FLTL practitioners and NLP specialists find a common framework to describe FL learning activities, the learner responses that they are expected to elicit, and the assessment procedures pursued. This methodology will be integrated in modern approaches to language teaching, and in particular it will be exemplified for the development of materials in instruction settings following a task-based approach. Moreover, it will make use of language as a crossroad between the FLTL and NLP research fields, and it will facilitate the specification of the computational requirements for NLP strategies, focusing on finite-state automaton techniques.

Our second goal is to introduce ICALL materials in real-world instruction settings by facilitating teachers the instruments for them to keep the control over the pedagogical design. To achieve this goal, we propose an ICALL activity authoring tool, as a means for FLTL practitioners to author FL learning activities including NLP-based automatic assessment without the need for them to be trained in NLP programming. Moreover, this authoring tool is tested with secondary school teachers in real-world instruction settings following a so-called blended approach to language learning, that is, one that combines face-to-face instruction with computer-based instruction.

With these two goals we aim to overcome the two shortcomings identified in this introduction, and, in the end, to promote a better connection between the FLTL and NLP research communities. Altogether, this should increase the amount of knowledge transferred from one community to the other.

1.6 Structure of the thesis

This thesis is structured in five parts: Part I, *Introduction*, Part II, *Background*, Part III, *ICALL tasks – Where FLTL meets NLP*, Part IV, *Enabling teachers to author ICALL activities*, and Part V, *Conclusions*.

Part I consists of this chapter and Chapter 2. Chapter 1 served the purpose of framing our research, motivating it from the different perspectives of FLTL, CALL and NLP, and presenting the goals of the thesis.

Chapter 2 overviews research on ICALL systems. We review the beginnings of ICALL research and the characteristics of the ICALL systems that have successfully integrated pedagogical and computational considerations in their design, implementation and evaluation for a sustained period of time. This chapter also reviews the current achievements and challenges in ICALL, among which we find the involvement of teachers and learners in the design and experimentation process. Accordingly, we review the role of the teacher in the process of authoring ICALL or, generally, CALL materials, and the research on the development of authoring tools.

Part II, *Background*, consists of two chapters, Chapters 3 and 4, which introduce the background concepts, on which the research presented in Parts III and IV is based.

Chapter 3 introduces Natural Language Processing as a field aiming to emulate human language understanding. In this chapter we introduce the concepts of domain and robustness, which determine to a great extent the success of NLP approaches in human tasks other than instruction and learning. We also introduce the computational techniques used to analyse text containing so-called ill-formed structures, that is, text that does not follow the standard and normative writing and linguistic rules of a given language – as is the case often for learner language. Particularly, we introduce the two main approaches to the analysis of ill-formed language – the so-called mal-rule approach and the constraint relaxation approach.

Chapter 4 presents the FLTL concepts with which this research aims to be compatible with. We introduce Communicative Language Teaching and Task-Based Language Teaching (TBLT) as the general approach in which we aim to integrate NLP-based automatic feedback generation. This introduction includes a review of the works that have tried to establish criteria for the qualification and the classification of FL tasks as communicative tasks. Moreover, we present reference work on the creation of TBLT-driven materials, both class materials and tests. Such works will provide us with frameworks for the formal pedagogic and linguistic characterisation of FL learning tasks.

Chapter 4 also includes a review of recent studies on the effectiveness and the use of feedback in settings where ICALL or CALL materials were used, studies through which we learn about feedback presentation techniques that have proven to significantly increase learning gain.

Part III of the thesis consists of five chapters and it describes the research we carried out to integrate pedagogical needs and computational capabilities along with the design and the development of ICALL tasks, meaning FL learning activities including NLP-based correction functionalities under a communicative approach to language learning, particularly Task-Based Language Learning. We propose methodological instruments to be used during the FL learning task design process. Every time we present such an instrument, we include a theoretical description and a subsequent exemplification in a concrete pedagogical setting, with the aim of reflecting the importance that both theoretical and practical aspects have.

Chapter 5 starts with some methodological considerations regarding the specificties of ICALL instruction settings, the type of interaction between learner and virtual tutors. This chapter also explores the ways in which the characteristics of ICALL instruction can determine the main elements that our object of our study, that is, the activity, the expected responses, the module for language analysis and the module for feedback generation.

Chapter 6 introduces a specific setting for the creation and validation of ICALL materials, a setting with particular pedagogical goals and pre-existing NLP tools for the analysis of learner language. This setting is characterised by following a TBLT approach to language instruction and relies on a modular NLP architecture for the analysis of learner language using shallow processing techniques and the mal-rule approach. We will provide the FL learning tasks on which we exemplify and validate the ICALL task development frameworks that we propose.

Chapter 7 introduces the framework we propose for the characterisation of ICALL tasks, which will allow us to describe them in terms of communicative goals, linguistic goals, assessment criteria, and linguistic characteristics of the language expected in the responses. The chapter exemplifies how this framework is applied to ICALL tasks of different natures, that is, tasks whose responses present different challenges in pedagogical and computational terms. The practical result of Chapter 7 is the characterisation of a series of ICALL tasks including a detailed characterisation in terms of thematic and linguistic contents expected and the assessment strategies required.

Chapter 8 presents the step in which the FLTL-driven characterisation of ICALL tasks is further elaborated to meet specific demands for NLP. In this chapter, we present two schemes for the transformation of the characterisation of ICALL tasks into formalised linguistic processing requirements for the modules responsible for the analysis of learner language and the generation of feedback in a particular ICALL setting. We will describe how the implementation for the rules to be included in the corresponding modules can be based on pedagogically informed design-based specifications.

To conclude Part III, Chapter 9 presents a study in which we analyse actual learner responses to ICALL tasks. The analysis compares the language elicited from learners by these tasks with the expectations stated in the pedagogical and computational characterisations presented in Chapters 7 and 8. Moreover, the analysis is an opportunity to quantify and evaluate the kind of variation that we find in learner responses in terms of thematic and linguistic contents. In addition, we discuss the effects that variation can have on the NLP complexity of particular ICALL tasks. Finally, the chapter shows how learner data can be used to enhance the performance and the coverage of domain-specific, that is, ICALL task specific, NLP strategies.

Part IV of the thesis consists of two chapters that present our research on the integration of ICALL materials in instruction settings with the proviso that teachers keep the control over the whole design, implementation, and use phases. To achieve this aim we propose a methodology that allows for the customisation of NLP-based feedback generation strategies to specific learning tasks. Moreover, we present an experiment with teachers and learners in real-world instruction settings in which the proposed methodology is implemented and co-evaluated.

Chapter 10 presents a methodology that allows us to surpass the intervention of an NLP developer in the design of ICALL materials. This chapter describes how the NLP-based feedback generation strategy presented in Part III can be adapted to become a customisable strategy. This is the prerequisite for teachers to author their own ICALL materials. The chapter presents a formal language defined to interface between NLP resource specifications and teacher-defined response requirements, an interface through which we obtain a teacher-specified list of correct and well-formed expected responses for each item in a FL learning activity. Additionally, we present an NLP-based technology by which teacher-provided responses are automatically expanded into a series of NLP models to handle a range of correct and incorrect responses. The whole adaptation of the NLP-based feedback generation strategy, the response specification methodology and the interface language are described and exemplified.

Chapter 11 presents an experiment through which the methodology to empower teachers with an ICALL authoring tool is implemented and evaluated. The chapter introduces the instruction settings and the software implementation in which the customisation strategy of NLP resources for NLP-based feedback generation is technically integrated. It presents the working process of the experiment, as well as the resulting products: the materials created by the participating teachers and the learning experiences of the respective learners. The chapter provides an evaluation of the experiment in terms of the characteristics of the generated ICALL activities and of the quality of the feedback provided to learners. Moreover, it presents a subjective evaluation from the perspective of teachers and learners, and a discussion.

Part V of the thesis includes Chapter 12, which presents the contributions of the thesis from the perspective of our two principal research goals and the future avenues for research envisaged in the short and the long run. It also summarises the most outstanding implications that our research findings can have on the more general fields of FLTL and NLP, respectively.

Chapter 2

The goals of the thesis within ICALL

This chapter frames the thesis within the field of Intelligent Computer-Assisted Language Learning (ICALL) particularly focusing on the notion of Intelligent Language Tutoring System (ILTS), and on the multidisciplinary nature of this research area. Within the ICALL field, we analyse the characteristics of those ILTSs with a long trajectory in the field that combined interdisciplinary research and practice, integrating views and approaches from the FLTL and NLP worlds. We review the literature concerned with current achievements and challenges in ICALL, particularly those related to the integration of NLP-enhanced feedback generation systems in instruction settings following a task-based approach to language teaching and learning. Additionally, we look at the few efforts made to develop authoring tools for the creation of FL learning activities including intelligent feedback, and also at the reason why this violates an important principle in CALL. Finally, we will revisit the goals of the thesis to contextualise them in current trends in ICALL research and to present how we expect our research to contribute to the field.

2.1 An overview of the research in ICALL

This section overviews the history of ICALL. We first describe the pioneering work aiming at the integration of NLP in foreign language learning, whose goal was the development of a so-called Intelligent Language Tutoring System – this work reflects well the essence of ICALL research. After that, we review 30 years of ICALL, and we highlight the most prominent research lines in the area, as well as the most important advances in it. We conclude with a description of the three ICALL systems that have been in use for a sustained period in real-world instruction settings, where we compare their pedagogical, design and technical features.

2.1.1 The beginning of ICALL

The first research in which NLP was used for the teaching and learning of a foreign language is the one by Weischedel, Voge, and James (1978). The authors developed a system for the assessment of reading comprehension responses from English learners of German as a foreign language for the first three weeks of a course for beginners. Their system can be considered the first ILTS ever.

From an instructional perspective, their goal was to "provide an additional tool to augment classroom instruction with comprehension and composition exercises", but never as a replacement for classroom instruction (Weischedel et al., 1978: p. 226–227). The actual implementation allowed for the automatic assessment of comprehension questions to two texts from two learning units of the textbook *Moderne Deutsche Sprachlehre* written by Duval et al. (1975).

From an NLP perspective, their system used what is commonly known as a symbolic approach to language processing and reasoning. First of all the system analysed learner responses with a parser implemented using Augmented Transition Networks. The linguistic parser was designed to process both well-formed and ill-formed language, and to handle them differently. After that, the syntactically analysed sentences were translated into a semantic formalism. Then, the semantically formalised sentences were used to check against a "world model" (the text's world) that contained the knowledge necessary to assess a set of correct responses, and a set of incorrect responses. The system added up a list of errors as the learner response went through the different analysis modules and finally this was translated into a list of feedback messages for the learner.

Weischedel et al. (1978: pp. 237–239) conclude their work with a set of interesting remarks on the limits and the advantages of their system, and particularly on the practicality of making this technology easy to use for FL teachers. In terms of the linguistic resources, the authors think the major limitation is the coverage of the lexicon and the syntactic component, as well as the adaptation of the parsers to learners in different levels of proficiency; however, they think this can be reasonably overcome. In contrast, they see as very problematic the limitations of the semantic model, since they think that the "texts that appear in foreign language textbooks very rapidly surpass the ability of artificial intelligence systems" (1978: p. 237).[1] Even simple sentences as "I almost always study alone" have to be changed to "I always study alone" – because of the difficulty of modelling the possible world that corresponds to adding *almost* to the sentence in terms of the scope of quantifiers over events. Nonetheless, alone extending and/or adapting the parser and the dictionary is a task that they consider time-consuming and too difficult for it to be executed by a language instructor (1978: p. 237).

In their view, the major advantage of their system is that it allows learners to freely use the language naturally in answering a question (1978: p. 238), something that was demanded at that time by researchers working in CALL (Nelson et al., 1976). This, complemented with a strategy allowing instructors to create semantic models for new activities, would allow for the integration of their approach in real-world instruction settings. However, this could not be done with without an "interesting burden", namely to itemize each fact implied in the world derived from

[1]Natural Language Processing is often referred to as Artificial Intelligence, though Artificial Intelligence includes various other subdisciplines, some of which, such as learner modelling, are also present in ICALL research. See Schulze (2008: p. 510).

the lesson's text (Weischedel et al., 1978: p. 238).

2.1.2 More than 30 years of ICALL

In their book *Errors and Intelligence in Computer-Assisted Language Learning: Parsers and Pedagogues*, Heift and Schulze (2007: p. 55–56) identify 119 NLP projects in CALL during the period between 1982 and 2004. Heift and Schulze record a total of 70 projects between the mid-1980s and the mid-1990s, and a total of 40 between the second half of the 1990s and the first of the 2000s.

Over the years NLP has been applied to language learning by developing (Heift and Schulze, 2007: Ch. 2, Schulze, 2010: p. 70–78):[2]

- So-called writer aid tools, which can help improve the quality of the learner's written production even though they are not designed to learn a language. Among these, some of them concentrate on the correction of FL learner errors in non-restricted domains and others in the correction of errors in limited domains. Most of them are specialised in the target audience. In this group we find among others Gamon et al. (2009)'s research on the correction of errors made by learners of English, or ICICLE (Michaud and McCoy, 2006), a system that supports the learning of written English to signers of American Sign Language as a first language. Rimrott and Heift (2008)'s research is also interesting since it analyses how learner-specific tools perform compared to tools developed for native speakers (of German).

- Systems concentrating on the teaching and the learning of specialised grammatical phenomena, such as the use of adjectival endings, the use of morphology and syntax in noun phrase elements, word order in sentences, the use of (clitic, zero) pronouns, and so on. In this group, we find *VERBCON* (Bailin, 1990) and TDTDT (Pijls et al., 1987), which focus on the appropriate usage of verbs with respect to a selection of linguistic phenomena; *SWIM* (Zock, 1992), focusing on the use of clitics in French; and ALICE (Cerri, 1989) focusing on the use of temporal constructions in Italian, French and English.

  In this group we can also include the only three systems that are still used in instruction settings today, *ROBO-SENSEI*, *E-Tutor* and *TAGARELA* (see Section 2.1.4) which are designed to be used in particular phases of a task-based approach to language learning with the goal to reinforce certain formal aspects of the learning experience (Schulze, 2010: p. 76–79).

- Systems focusing on the teaching of specific communicative competences to language learners. Among them we find a system to chat with the computer about one's family, another about buying food in the market, or role-play activities to play spies or private investigators. In this group Schulze (2010: p. 70–73) includes *FAMILIA*, a system to chat about one's family that pays attention to

[2]This list excludes applications of NLP for automatic scoring of learner essays, as well as tools for the automatic annotation of learner corpora because strictly speaking they are not applications for learners to learn a language.

particular verb complement combinations (Weizenbaum, 1976); *Spion* (Sanders and Sanders, 1995) and *Herr Komissar* (DeSmedt, 1995), two systems that respectively use the spies and the private investigator domains to engage learners in a game-like conversation; the work by Menzel and Schröder (1999), where learners state utterances related to a graphical market scenario; and *FLUENT-1* (Hamburger and Hashim, 1992) and *FLUENT-2* (Schoelles and Hamburger, 1996), a graphical system in which learners could move objects in a particular micro-world (a bathroom) per request.

- Reading support tools, such as dictionaries or morphological analysers hyperlinked to reading texts, or links from the reading text to concordancers as a means to learn more about the usage of selected words. In this group we find *GLOSSER RuG* (Nerbonne et al., 1998; Roosmaa and Prószéky, 1998) and *ELDIT* (Knapp, 2004), two slightly different tools that assist learners in reading activities and vocabulary acquisition; and *QucikAssist* (Wood, 2009), a tool that allows learners to obtain linguistic and encyclopaedic information related to words in a text by clicking on them.

According to Heift and Schulze (Heift and Schulze, 2007: Ch. 2, Schulze, 2010: p. 70–72), ICALL systems with smaller coverage and less ambitious goals are the ones that commonly went beyond the prototype and reached the language learner. Examples of such systems are *Spion* (Sanders and Sanders, 1995), *Herr Komissar* (DeSmedt, 1995), *GLOSSER RuG* (Nerbonne et al., 1998; Roosmaa and Prószéky, 1998), *ELDIT* (Knapp, 2004), *ROBO-SENSEI* (Nagata, 2010), *ETutor* (Heift, 2010b) and *TAGARELA* (Amaral, 2007; Ziai, 2009; Amaral et al., 2011; Amaral and Meurers, 2011). More ambitious projects have yielded interesting results, but they usually end up not being used by learners: Two interesting examples are *Textana* (Schulze and Hamel, 1998; Schulze, 1998, 1999, 2001, 2003) and *Freetext* (L'Haire and Faltin, 2003; Granger, 2003; L'Haire, 2004).

Core research issues and influences from other disciplines

Over these 30 years of ICALL, different kinds of problems were approached, and different solutions attempted or adopted. The three issues most frequently tackled over the years are (i) the analysis of learner language, (ii) the appropriate strategies for the provision of feedback, and (iii) the adaptation of feedback to learners with different learning profiles and styles. We focus on the analysis of learner language in Chapter 3, where we introduce the key issues in Natural Language Processing for the purposes of this thesis.

As for the other two topics, Feedback and Student Modelling, we introduce the aspects that were most significant in ICALL according to Heift and Schulze (2007: Chs. 3 and 4). As for Feedback, it is generally understood as corrective feedback, and the main challenges in ICALL are to make feedback clear, comprehensible, as profitable as possible, and, of course, pedagogically grounded (Heift and Schulze, 2007: pp. 115–116). Heift and Schulze argue that (I)CALL research addressing the topic of feedback benefits from considering the general points of view of human-computer interaction (HCI), learning theories, second language acquisition theories, and formal grammar (2007: p. 116). As we will see in Section 4.5, several studies have analysed how language learning is affected by the number of feedback messages, the wording, the inclusion of graphical highlighting, the grouping or filtering of corrective feedback depending on the relevance and the nature of the errors, the steps in which learners access different levels of feedback and the corresponding cognitive load – see also (Garrett, 1987; Pujolà, 2001; Nagata, 1993, 1995, 1996).

As for Student Modelling, ICALL research is influenced by the research in user and student modelling in Artificial Intelligence. This influence contributed to a better understanding of how user information can be stored, what information is required in order for the student model to communicate with other modules in the system, and how the characterisation of the student as a learner can be updated over time according to his or her progress. ICALL practice has brought up topics discussed in student modelling such as the criteria to balance the weight of learner performance according to the learner's developmental stage, the goals of the activity, the frequency of a particular error, and so on. Student Modelling falls out the scope of this thesis and will not be addressed in the following chapters. Further readings on this topic: Matthews (1992), Bull et al. (1995), Heift and Schulze (2007), Amaral and Meurers (2008).

Sustained research and development

An important aspect of ICALL research is the sustained use and development of ICALL systems. CALL systems, in general, are systems that are, should be, permanently improved and adapted to the teacher and learner needs. In fact, those systems that present a sustained use and progress over the years are the ones that managed to be successfully integrated in real-world instruction settings – and they deserve particular attention (Levy, 1997: p. 13–14, Heift and Schulze, 2007: p. 9).

As for ICALL systems, researchers (Schulze, 2010: p.76–77, Amaral and Meurers, 2011: p. 3) agree that, there are only three systems that have been used for a sustained period of time in real-world instructions settings: *ROBO-SENSEI*, *ETutor* and *TAGARELA*. The reasons for this are fundamentally related to the difficulty of combining research and development (including use with learners), and the complexity of putting and keeping together cross-disciplinary teams (Antoniadis et al., 2004; Heift and Schulze, 2007; Schulze, 2008; Amaral and Meurers, 2011). Thus, rephrasing our intial reference to Heift and Schulze (2007: p. 9), we claim that ICALL is a field of investigation that demands *multidisciplinary research* and *sustained development*.

2.1.3 The essence of an ILTS

According to Levy and Stockwell (2006: p. 22), an Intelligent Language Tutoring System (ILTS) can be defined as "a computer program [that] analyses and evaluates an individual learner's response to a question, and provides feedback on it". But ILTSs do not necessarily imply the processing of learner responses (reactions) with Natural Language Processing: responses by mouse actions (clicking, drag-and-drop, circle), reaction time (time spent in or frequency of use of certain resources), or, at least for research purposes, eye-tracking are all possible ways through which interaction can take place. There might be good reasons for expecting learner interactions based on these types of responses, for instance, in order to keep low the cognitive demands on the learner side, as in Pujolà (2001: p. 83), discussed in Section 4.5. However, in using an ILTS that requires learners to provide language-mediated responses, NLP is compulsory. There are systems that instead of using NLP analysis use string character comparison techniques, or lists of possible responses for which a feedback is considered, but this is clearly an insufficient strategy – see also Nagata (2009) for a good example of how useless such strategies can be even for relatively simple language learning activities.

2.1.3.1 Architecture and functionalities of an ILTS

According to Burns and Capps (1988: Ch. 1) Intelligent Tutoring Systems (ITS) consist of an expert module containing the domain knowledge, a student module determining what the student knows, and an instructor module identifying the "deficiencies" in knowledge to focus on what will determine the strategies for presenting the knowledge to be acquired. Moreover, they add, "the instructional environment and human-computer interface channel tutorial communication" (1988: p. 2). This modular organisation is also followed by researchers in ICALL (Heift and Schulze, 2007; Amaral, 2007).

In essence, an ILTS has to provide the necessary functionalities for the tutor-learner interaction to take place: the tutor offering a set of activities, to which learners react; and the tutor providing feedback to learner (re)actions. Figure 2.1 reflects how the architecture of an ILTS is according to Amaral (2007: p. 85).

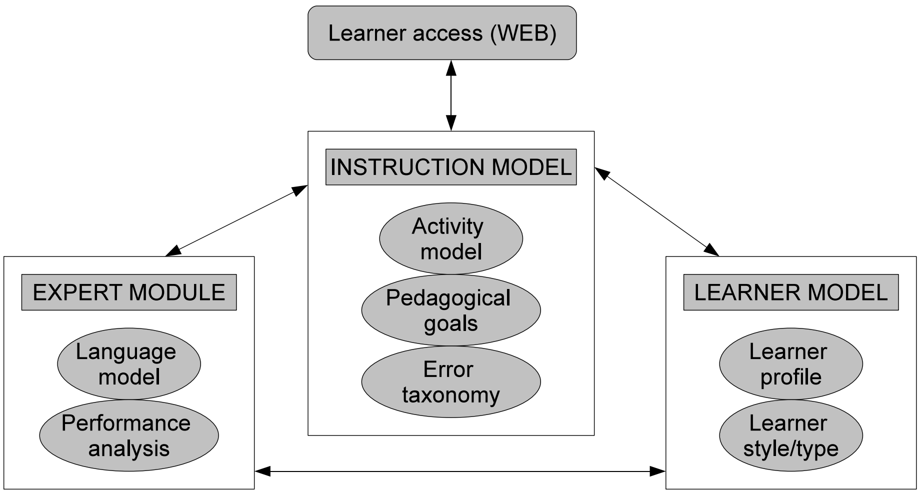


Figure 2.1: General architecture of an Intelligent Language Tutoring System – simplification of the one proposed in Amaral (2007: p. 85).

The functionalities that result from the interrelation between the modules in Figure 2.1 are:

1. Presentation of instruction activities: The instruction model is responsible for the content presented to learners. It can also inform the learner and the ex-

pert module by requiring from them functionalities relevant for the instruction context.

2. Learner activity: The learner model centralises information about learners coming from the other two modules, but also includes information about the personal characteristics of the learner. The instruction model and the learner model together monitor learner progress within each proposed activity.

3. Tutoring: The expert module is responsible for providing the learner with appropriate feedback related to his or her performance, and particularly to his or her language production. The (virtual) intelligent tutor itself will be based on the results provided by an NLP-based algorithm together with an algorithm that enables the system to interpret those domain-related data in the context of a particular instruction goal.

These functionalities are present in virtually any software trying to teach a foreign language through activities requiring learner responses. For instance, in Weischedel et al. (1978) we had a syntactic model, in that case the domain model, and a world model, in that case an activity model. There was also an error taxonomy explicitly defined and used in both the syntactic model and the domain model. And one could argue that there was even a static student model, since the software had been conceived for beginner learners of German with English as a first language.

2.1.4 ICALL systems in use

In this section, we describe the most relevant aspects of the three ICALL systems that are fully integrated into real-life foreign language learning programmes at the university level. First, we provide some basic information on each of the systems: language taught, approximate starting date according to research publications, the levels of proficiency for which they are conceived, and some general comments on their use. Then we compare how theses three systems were conceived, developed or tested with regard to a series of issues particularly relevant for ICALL systems: instruction context, pedagogical orientation, design principles, system architecture, language processing strategy, feedback generation and learner modelling.

Basic system information

ROBO-SENSEI (Nagata, 2004) was initially called Nihongo-CALI (Nagata, 1995), and over a period of time was also known as BANZAI (Nagata, 1997b). The work around this system began with Nagata's thesis (1992), and several papers on its design, its development and its use were published over the years – the most recent being Nagata (2010). The system was developed for teaching Japanese to English speakers at the university level. Although it was mainly used with beginner courses, the author claims that its NLP analysis engine "can process all grammatical constructions introduced in a standard Japanese curriculum from beginning through advanced levels" (Nagata, 2002: p. 584). This system is currently commercialised through a publishing house – see http://www.cheng-tsui.com/store/products/robosensei.

ETutor (Heift, 2004) is the name of a system previously known as the German Tutor (Heift and Nicholson, 2001). ETutor is an ICALL system whose development started with Heift's thesis (1998), and it is still further maintained and improved Heift (2010b). It was developed for the teaching of German to beginners at the university level in Canada (Simon Fraser University). It has been used there for a long time, and it is available for use at other universities for a semester-valid user fee (Nina Vyatkina, p. c.) – see also http://www.etutor.org.

TAGARELA, which stands for Teaching Aid for Grammatical Awareness, Recognition and Enhancement of Linguistic Abilities (Amaral, 2007: p. 50, Amaral and Meurers, 2006), is a system developed and piloted in 2006 and first used in regular Portuguese teaching courses in 2007. It has been used at The Ohio State University and is currently being used at the University of Massachusetts Amherst, and the authors are extending its contents and improving its technical functionalities (Amaral et al., 2011). The system is aimed at instructing learners of Portuguese in their first courses at the university level and specifically at students in Individualised Instruction Language Programs in a North American context (Amaral, 2007: pp. 49–53) – see also http://purl.org/icall/tagarela.

Detailed feature comparison

Instruction context The three systems (and their predecessors) have been and are being used in combination with face-to-face instruction (Nagata, 1993, 1995, 1997b; Heift and Nicholson, 2001, Reeder, Heift et al. 2001) and as a complement to standard textbooks (Nagata, 2010; Heift, 2004, 2010b). ETutor is nowadays adapted to accompany the first three German courses at the university level when using the book *Deutsch: na klar!* (Heift, 2010b; Di Donato et al., 2004; Sanders, 2012), while ROBO-SENSEI is evolving into an independent textbook for Japanese (Nagata, 2010). As for TAGARELA, it was initially designed be used as an intelligent electronic workbook integrated in the Portuguese Individualized Instruction Program at The Ohio State University (Amaral, 2007: pp. 51–64). TAGARELA is also used in combination with face-to-face instruction (Amaral, 2007: pp. 132–133), and is used in distance education courses at the University of Massachusetts Amherst.

As for **curriculum embedding**, the three systems are being used in communicative instruction settings at the university level in combination with standard instruction textbooks or other electronic materials – Nagata (2002: p. 583), Heift (2001a,b, 2010b), and Amaral (2007).

Pedagogical orientation The three systems explicitly follow SLA and FLT theories where corrective feedback and focus on form are considered, and are always embedded in settings following a communicative approach (Nagata, 1993, 1995, 1998; Heift, 2001a, 2003; Amaral, 2007: pp. 46–48). Moreover, Nagata has investigated the benefits of explicit inductive feedback – that is, feedback driven to help learners figure out the linguistic rules underlying certain structures or sentences – and language production practice (1997b), and her system is designed to produce this kind of feedback. As for TAGARELA, according to

Amaral it "was designed to help fill one common gap of language instruction at the university level in the United States: the lack of personalized feedback students receive on their language production because of the small amounts of time instructors can spend with each individual student" (2007: p. 50, but also the whole of Ch. 3). All three systems are consistent with the corresponding North American syllabi.

As for their **pedagogical goal**, both ROBO-SENSEI and ETutor aim at fostering the acquisition of grammar and vocabulary competences of beginner to intermediate levels of the second language. ROBO-SENSEI includes linguistic aspects such as sentence particle usage, verb inflection, auxiliary verbs, and passive voice – see Nagata (1995: p. 51), Nagata (2002: p. 598) and Nagata (2010: p. 461) –, and ETutor includes aspects such as noun phrase agreement, subject agreement, auxiliary verbs, and use of punctuation – see Heift (2003: pp. 542–545). TAGARELA, in addition to being aimed at the acquisition of certain grammar and vocabulary competences, also aimed at fostering listening and reading comprehension skills (Amaral, 2007: pp. 70–71). It covers spelling and grammar errors and provides information about the semantic appropriateness of student production on the basis of shallow content analysis (Amaral, 2007: pp. 90–91, Amaral and Meurers, 2008: pp. 321–322).

As for the **activity types** offered, all three systems offer exercises where relatively-free (pedagogically constrained) short answers are required. In ROBO-SENSEI the learner is required to write a sentence following specific instructions on what to write (Nagata, 1997b: p. 518) – since the instructions are provided in English they can be considered translation exercises, as pointed out in Amaral and Meurers (2011: pp. 8–9). Its forthcoming version will include activities where listening and reading comprehension activities are found, as well as character writing activities (Nagata, 2010). ETutor included from the beginning dictation, build a phrase, which word is different, word order practice, fill-in-the-blank, and build-a-sentence activities (Heift, 2001b), and later on has incorporated reading comprehension, listening comprehension and short essays (Heift, 2010b).[3] As for TAGARELA, it includes listening and reading comprehension, descriptions, vocabulary, rephrasing, and fill-in-the-blank (Amaral, 2007: pp. 64–79).

Design principles Underlying ROBO-SENSEI's design choices there is the need for an interactive system that can provide an immediate response, that enhances the student-textbook interaction and favours self-paced learning, that provides specific and linguistically principled feedback to correct or incorrect input, and that encourages the development of production skills (Nagata, 2010: p. 461 and 463). Similar criteria are provided by Heift, where she mentions the emulation of "learner-teacher interaction" (Heift, 2010b: p. 445). The authors of the three systems argue for sophisticated answer-processing tools because it is not feasible to anticipate every possible mistake made by learners. Quite appealing in this sense is the procedure followed by Amaral (2007). To better define and learn about the language teaching context in which the system is to be integrated he

[3]Short essays are corrected manually by teachers.

interviews teachers of Spanish and Portuguese as to their real-life needs (Amaral, 2007: pp. 53–64). In all cases, NLP limitations were taken into account and specifically tackled by appropriately restricting the language required through careful activity design (see Amaral and Meurers, 2011).

System architecture All systems follow similar processing architectures, but the most thoroughly discussed one is TAGARELA's Amaral et al. (2011). It consists of six modules: the interface; the analysis manager, which allows to configure and select the appropriate NLP tools for each learning activity; the language processing module; the feedback manager; the learner model; and the instruction model (Amaral, 2007: pp. 84–85). These modules or corresponding functionalities are present in the other two systems too (Heift and Nicholson, 2001; Heift, 2003; Nagata, 1997b, 2002). All systems use client/server architectures and include *ad hoc* learning management functionalities. The programming languages they use are JAVA, Prolog, LISP, cT, and Python.

Language processing strategy The **language processing architecture** in the three systems evolved – one should say evolves – over the years (Nagata, 1995, 1997b, 2002; Heift and Nicholson, 2001; Heift, 2003; Amaral, 2007; Ziai, 2009). They all present an architecture consisting of a processing pipeline including word segmentation, spell checking, lexicon look-up, part-of-speech tagging, partial syntactic parsing, specific grammatical or language use and error detection (e.g., agreement checker). TAGARELA includes simple semantic (or content) checking on the basis of shallow linguistic information (Amaral, 2007: p. 95–110). ETutor and TAGARELA use a combination of mal-rule detection techniques and relaxation techniques, ROBO-SENSEI relies exclusively on mal-rule techniques often intermingled with the language analysis process. All three systems opt for keeping things simple whenever possible, so string matching techniques are also strategically used to reduce the response time of the system.

TAGARELA was recently re-implemented using UIMA[4] as an underlying software platform, which results in a more flexible, modular architecture that allows for the specification of the processing modules on the basis of the required NLP tasks (Ziai, 2009: pp. 16–18). UIMA is based on a data structure that allows for the flexible combination of analysis features which can become as complex as required – not only linguistic features determined by the NLP, but also any other type of information that can be determined by the learner or instruction models.

As for **linguistic knowledge representation**, all systems provide theoretically informed linguistic structures, but only ETutor uses a sophisticated formalism for the representation of linguistic information. Heift and Nicholson (Heift and Nicholson, 2001; Heift, 2003) use Head-driven Phrase Structure Grammar (HPSG), where linguistic information is formally represented as feature structures which encode partial descriptions of a linguistic sign following a lexicalist approach (Pollard and Sag, 1994). Heift's approach to the analysis of ill-formed

[4] http://uima.apache.org/

input is to relax the constraints imposed by features such as gender, number or case. Ill-formed constructions are allowed, but marked as incorrect. This way, a combination of incompatible linguistic features is allowed and recorded in so-called descriptors, whose information is percolated to their respective heads. This information can later on be used to provide information about both correct and incorrect learner production. TAGARELA's newer version uses UIMA types, which "are equivalent to typed feature structures used in formalisms such as HPSG" (Ziai, 2009: p. 44).

As for **computing algorithms**, all systems use standard finite-state automata techniques for certain modules (word segmentator, disambiguator, agreement checker, content checker), and some version of the Damerau-Levenshtein algorithm for spell checking, except for Japanese, where a mal-rule approach to certain spell checking error types is used. As for the syntactic tree-building algorithms, ROBO-SENSEI uses a Generalised Left-to-right Rightmost (GLR) parser (Nagata, 2009: p. 566), and TAGARELA's recent version uses an implementation of the Cocke-Younger-Kasami (CYK) algorithm (Ziai, 2009: pp. 49–58). ETutor uses genetic algorithms for the detection of incorrect word orders (Heift and Nicholson, 2001).

Feedback generation In the three systems feedback generation is a two-step process: (i) error diagnosis and (ii) the actual presentation of feedback messages to the learner (henceforth feedback presentation). In the first versions of ROBO-SENSEI, **error diagnosis** was encoded in the parser's rules following a mal-rule approach (Nagata, 1995). As described in Nagata (1997b), the system checks on a surface level the learner's answer and, after an initial filtering based on heuristics, it selects and analyses the answer stored in the system with a higher proximity to a target answer encoded in the system as part of the activity. This target answer is used to compute all the error diagnosis operations. In the latest version of the processing architecture, error diagnosis modules are separated from analysis modules but interleaved in the processing sequence (Nagata, 2002: pp. 590–592). ETutor performs error diagnosis during the parsing itself, by using relaxation techniques. Since the parser incorporates the ability to handle ill-formed input, no added procedure is needed for the detection of sentence-level errors. Errors related to extra or missing words, or wrong word choices are handled using other mechanisms ranging from string-based pattern matching to more complex rule-based or statistics-based algorithms (Heift and Nicholson, 2001). The error diagnosis amounts to interpreting the analysis provided by the NLP tools in the context defined by the activity model and the learner model. TAGARELA follows a strategy in which each module contributes to the analysis and evaluation of the learner response with its specific functionalities: the resulting analysis is a combination of the different analyses, sometimes overlapping, provided by the different modules (Ziai, 2009: p. 39).

As for **feedback presentation**, ETutor and TAGARELA follow a similar strategy. A specific and externalised feedback module collects all errors detected by the diagnosis module and orders them according to specific criteria, which can be configured – e.g., errors related to the specific goal of a learning activity are prioritized (Heift, 2003: pp. 543–544, Ziai, 2009: pp. 40–41). These two systems present only one feedback message at a time. ETutor is in this respect a pioneering system in that its authors worked out a strategy to adapt the feedback to the learner's level; depending on learner performance, different feedback messages reflect different degrees of explicitness. In contrast ROBO-SENSEI's feedback results from a mapping of all the collected error codes to feedback messages. Feedback messages are grouped into categories of an internally defined typology, which includes classes such as missing word, particle error or predicate error (Nagata, 2002: p. 592).

Learner modelling ETutor (Heift and Nicholson, 2001; Heift, 2003) and TAGARELA (Ziai, 2009: pp. 34–35) include learner modelling capabilities. Both systems collect and maintain information about learners profiles and behavior. In Heift's words this allows for "modulation of instructional feedback" and "assessment and remediation" (2003: p. 541). In both cases the modelling is based on a network structure whose nodes correspond to grammar skills (grammar phenomena handled by the parser) for which the student is internally penalized or rewarded (node scores). This allows for a classification of students into beginner, intermediate and advanced learners. Every time a learner receives feedback on a specific grammar skill, this feedback is made more or less explicit according to his or her recorded performance.

2.2 Task design and automatic language processing

This section consists of two subsections. The first one reviews the efforts made in ICALL research to adapt to modern approaches to language teaching, particularly to task-based language teaching. The second one describes the only work that, to our knowledge, has argued for the need to characterise foreign language learning activities as a means to identify those that are both pedagogically meaningful and computationally feasible.

2.2.1 The pedagogical purpose as a driver of ICALL research

As we pointed out in Section 1.4, the mutual disenchantment between the NLP and the FLTL and CALL communities is related to the low proportion of truly cross-disciplinary research in ICALL. The need for pedagogically principled design of ICALL is stressed by several researchers in the literature, among them Schulze (2008), Heift (2010a), Nagata (2010), Schulze (2010), and Amaral and Meurers (2011). According to Schulze (2008), research in CALL including NLP attempted to:

1. prove by concept that ICALL systems can be used for practising very specific communicative skills

2. integrate focus-on-form assessment as pre-task activities, or as during-task and post-task support in task-based language instruction
3. assist learners by providing them with appropriate and adaptive feedback to controlled production activities
4. expand the coverage of the employed dictionaries and grammars to cover language from many different domains
5. concentrate on the selection of specific linguistic phenomena which are both relevant for FLTL and feasible in NLP
6. embed the NLP technology in games and virtual worlds
7. including learner modelling components in their architectures
8. providing several types of interaction modes with the learner in addition to corrective feedback, such as intelligent chat-bots or context-sensitive assistance for reading activities

Schulze (2010) analyses the efforts of ICALL researchers to make their research compatible with communicative language teaching. He explains how small-scale approaches to very restricted domains provided the context to implement ICALL systems that focused on the development of communicative competence of language learners (Schulze, 2010: p. 70–79). He distinguishes systems used in communicative tasks, systems used for during-task and post-task support, and systems used for pre-task activities. Examples and descriptions of ICALL studies under this threefold classification can be found in Schulze (2010: p. 70–79).

The research reviewed by Schulze (2010: p. 70–79) illustrates the feasibility of developing pedagogically principled and meaningful ICALL activities. However, the need for more of these pedagogically driven approaches still is a concern when focusing on the fundamental aspects of ICALL (Antoniadis et al., 2004; Heift and Schulze, 2007; Schulze, 2010; Amaral and Meurers, 2011).

In this respect, it seems that the theoretical and practical principles of FLTL and SLA should be present in ICALL design from the beginning along with the principles and the limits of NLP-based language processing. This latter idea is present in Heift (2010a: pp. 445–446) and Schulze (2010: p. 68), who rely on Colpaert's notion of *cyclical design* as the approach that most profitably can help bridge the gap between language pedagogy and technology. In this view, the cycle of design is a process that goes through design, development, implementation and evaluation as the only way to increase "the likelihood of a successful development outcome" (Schulze, 2010: p. 68).

2.2.2 From the focus on form to the focus on meaning

An issue that is relevant for ICALL in order to be more useful in communicative language teaching is the ability to assess learner responses in terms of meaning, that is, in terms of the contents, not only the form – the language. Bailey and Meurers

(2008) propose to delimit ICALL activities within the spectrum of FL learning activities as those that are pedagogically meaningful and computationally feasible, an area they call the **viable processing ground**.

Figure 2.2 is an abstract representation of the spectrum of FL learning activities as proposed by Bailey and Meurers. In one of the extremes we find activities that elicit tightly restricted responses requiring minimal analysis to be assessed. In the other extreme, we find activities that elicit unrestricted responses requiring extensive form and content analysis to be assessed. The viable processing ground lies between the extremes: It contains FL learning activities that are common in learning situations, that combine elements of comprehension and production, and are meaningful and suitable for an ICALL setting. From a form-based NLP perspective the responses to these activities will exhibit linguistic variation on lexical, morphological, syntactic and semantic levels, but the intended contents of the responses are predictable.

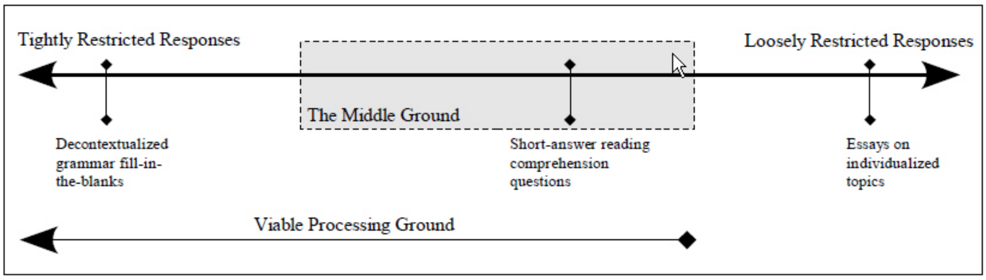


Figure 2.2: The viable processing ground (Bailey and Meurers, 2008: p. 108).

In Bailey and Meurers (2009: p. 4), the authors elaborate further the idea of assessing meaning in ICALL, and highlight the importance of "careful activity design" as a key to controlling variation in learner responses. The authors suggest three criteria to determine whether a FL learning activity is suitable for automatic processing: (i) the expected response variation, (ii) the availability of a gold standard, and (iii) the assessment criteria to evaluate the activity – mainly related to having a focus on meaning or a focus on form.

With regard to the expected response variation Bailey and Meurers (2009: p. 9) allude to two main characteristics of the activity: one of them is related to the response's length and the way it is correlated with the complexity of automatically analysing learner responses. The second one is related to the way in which the activity instructions constrain the elicited response. The latter, they suggest, can be achieved by constraining the response explicitly or implicitly, since instructions might determine the range of variation in leaner responses. According to them, linguistic variation can happen at many levels, and the changes at any of these levels might require changes in the remaining levels, such as a change in the structure of a sentence might require changes in the morphological endings of verbs. Their conclusion is that how the instructions constrain the learner response critically determines the suitability of FL learning activities to become ICALL activities.

As for the availability of a gold standard, Bailey and Meurers (2009: p. 10) emphasise that not only is it necessary that correct or acceptable responses to the activity can be identified, but also it must be possible to establish a set of responses that capture the essential contents (meaning) of the responses, so that the characterisation of acceptable variation can be established when performing meaning-based assessment. As for the assessment criteria, Bailey and Meurers (2009: pp. 10–11) state that it is important to define whether assessment of learner responses has to focus on the learner's grammatical competence or rather on its ability to use language to communicate in a task setting. Bailey and Meurers (2009: pp. 11–12) provide two examples of activities in the viable processing ground: reading comprehension activities that elicit short answers through specific questions, and a particular type of summarisation activities.

2.3 Tools for teachers to author FL activities

In this section we describe the only work that to our knowledge has pursued the development of an authoring tool for ICALL materials; we also review two other studies that investigate the automatic generation of certain types of activities including NLP-based feedback generation. Before we explain that, we contextualise in the broader frame of CALL the influence and the importance of authoring tools as a guarantee for teachers to keep control over content definition and content presentation.

2.3.1 Teacher control over CALL materials

The 1980s saw the beginnings of a closer involvement of the teachers in the conceptualisation of CALL materials (Levy, 1997: p. 43). The importance of the authoring of CALL materials can be traced in (Levy, 1997: ch. 2), and, certainly, there is a number of existing CALL authoring tools, such as Hot Potatoes (Arneil and Holmes, 1999), Dasher, HyperCard, TookBook, or WinCALIS. However, there is almost no research on the development of tools to author CALL activities with NLP-based automatic feedback, except for Toole and Heift (2002b).[5]

Interestingly, though, researchers in ICALL seem to be conscious of the importance of making the teacher autonomous in the development of ICALL activities. Already Weischedel et al. (1978: p. 239), in the plans for future work, speculate on the possibility that teachers specify the responses to the questions and the system automatically generates the necessary information for the NLP to work:

> [The instructor] would type in the questions to be asked; the [ICALL] tutor could automatically compute the minimal necessary information for an answer in a way similar to the computation of presuppositions. [...] [I]t should be technically possible for an instructor to write a lesson without becoming a systems programmer.

[5]There is research that included a teacher module for the generation of translation exercises, but generating translation exercises is different from generating foreign language learning exercises; in translation exercises the freedom the student has to respond to the demands of the activity is even more constrained (see Chen and Tokuda, 2003, and Rösener, 2009).

Similar opinions are found in Nagata (1997b: pp. 516–517), Toole and Heift (2002a,b), or Antoniadis et al. (2004: p. 19).

CALL research emphasises that CALL authoring software is key in the acceptance of CALL materials among teachers. This is explained by the fact that authoring tools offer teachers flexibility in "the presentational and instructional formats", i.e., they provide teachers with control not only over content, but also over "the way the content is presented" Levy (1997: p. 19). Levy and Stockwell argue that, compared to third-party materials, teacher materials are more easily integrated in the course programme and are usually designed "with the needs and resources of the individual learner in mind" (2006: p. 11-12).

Heift and Schulze (2007: p. 63) add two more arguments for the development authoring tools. First, both CALL and ICALL systems demand significant computational expertise as well as subject-domain expertise to design appropriate and contextualised learning materials. Therefore, teachers cannot be expected to undertake this on their own. Second, the task of providing updated, authentic, and relevant material is very time-consuming. Thus, this is probably a task that cannot be undertaken by NLP or CALL specialists and would benefit from the possibility of reusing and recycling materials.

2.3.2 Tutor Assistant

Toole and Heift (2002b) developed *Tutor Assistant*, an authoring tool for an ILTS for English with the goals of reducing the costs in time and expertise required to create ICALL activities. The system allowed for the generation of form-focused activities: build-a-sentence, fill-in-the-blank, and drag-and-drop. The resulting activities could be used in a web-based content manager that was an experimental counterpart of ETutor for English.

During the activity creation process, teachers decided what activity they wanted to create and provided with two types of information (2002b: pp. 378–379): First, they gave the input (the prompt) for the learner, a list of words to build a sentence, a list of word or sentence pairs to be matched, or a list of sentences with blanks. Teachers could also add some text with instructions besides or instead of a particular word, such as *(pronoun) / wash / (determiner) car / (reflexive)*. Second, they provided with all the possible correct answers to that activity in a pre-established manner.

When the activity was compiled, the system checked that the words in the activity were in the system's dictionary: If a word was missing the system included the functionality to add new words and a minimum of morphosyntactic information. The system also checked that all the responses typed in by the teacher did not contain any spelling or grammar errors, and that they corresponded with the activities instructions – that is, that they actually contained the words given in the instructions. This process freed the teacher from the need to implement the expert module (in the ILTS), and the task became a foreign language activity design task, a task teachers are familiar with.

Toole and Heift used the tool with a group of four teachers with different backgrounds in terms of teaching experience and in terms of computer literacy (2002b:

p. 380–381). Teachers were asked to produce activities for beginners of EFL on different topics, and they were asked to use Task Generator for 90 minutes.

From their study, Toole and Heift (2002b: p. 383–385) concluded that, in such an environment, exercise development-time ratios were "within the scope of a typical language teacher", which they assumed to be a maximum of five hours per week for a teacher with no expertise in developing an ILTS. They also concluded that the quality of the resulting activities was suitable for use in an intelligent language tutor. However, they also found that independent of their teaching experience and their digital literacy all users made some spelling and grammar errors. According to the authors, this supported the decision to include a validation component that checked for problems such as having typed in responses with typos, or responses that included more or fewer words than those included in the input data for the learner. According to interviews they conducted with the participating teachers, generally, it was not easy for them to create the activities, though the hardest work was to think of challenging exercises. Moreover, teachers liked being able to build activities for a complex system such as an ILTS, which, in addition, was very convenient to do on the web.

2.3.3 Automatic generation of ICALL activities

There is a similar research line that was also started by Toole and Heift (2002a) whose goal is the generation of ICALL activities by taking as input authentic texts. In this context, teachers would not be able to author activities including questions, but they would be able to decide on which particular text, say, a fill-the-gap or a multiple choice questionnaire should be generated. Toole and Heift (2002a) used NLP techniques to analyse texts written by native speakers for the automatic generation of FL learning activities. They developed a system that supported the creation of build-a-sentence, fill-in-the-blank, and drag-and-drop. They developed it for English and though the evaluation of the system yielded reasonable quantitative and qualitative results, they do not report on using the created materials with learners.

Meurers et al. (2010) developed a web-based service for the generation of activities such as fill-in-the-blank, identification of lexical or syntactic items in a text, and simple multiple choice activities. It also allows for the highlighting with colours of particular words in a text, words selected according to pedagogical criteria. The system was initially developed for English and was later extended to German and Spanish. It is offered as a web-based tool or as a Mozilla FireFox plug-in. Currently, they have plans to use it for experimental research in FLTL and SLA studies concerned with the use of consciousness raising strategies drawing the learner's attention to specific language properties – so-called input enhancement strategies (Sharwood Smith, 1993: p. 176).

2.4 Revisiting the goals of this thesis

The current challenges of ICALL are:

- The combination of complex and multidisciplinary requirements of the pedagogical and computational concepts during the design phase of the ILTS. These requirements need to be revisited during different cycles of development and use (Nagata, 2010; Heift, 2010a; Schulze, 2010).
- The meaningful integration of ICALL into current foreign language teaching and learning practice (Schulze, 2010; Amaral and Meurers, 2011).
- The exploitation of the data produced by learners, actual learner production, but also learner activity tracking data, during the the use of ICALL systems. Use the collected data to deepen our knowledge on fundamental research and practice questions such as the characterisation of learner interlanguage, or the use of learner performance as a measure for task complexity and as a criterion to inform changes in activity design. And, finally, to assess the usefulness of learning assistance tools such as different help options and feedbacks of different nature (Heift, 2010a; Schulze, 2010; Amaral and Meurers, 2011).

This thesis's goals, presented in Chapter 1, focus on the first two of the challenges:

- Our first goal is to propose a methodology that helps both FLTL practitioners and NLP specialists find a common framework to describe FL learning activities, the responses that they are expected to elicit, and the assessment procedures. This will be a contribution to further characterise the viable processing ground including insights from FLTL, CALL and NLP.
- Our second goal is to design and evaluate an infrastructure for FLTL practitioners to author and employ FL learning activities including NLP-based automatic feedback generation without the need of programming abilities. Crucially, the responses to the activities authored will be limited in length, but they will be more complex responses to those required in build-a-sentence or fill-the-gap activities. This will pursue the goal to foster the meaningful integration of ICALL in real-world instruction settings, as well as to facilitate learner individual work using computer-assisted instruction without reducing teacher control or autonomy.

In the following two sections we develop further the goals of the thesis, which, as we will see, presuppose a tight relationship and information transfer between FLTL, CALL and NLP, at the research level and at the practical level.

2.4.1 The feasibility of ICALL

Our first goal can be worded as characterising **the feasibility of ICALL**. This goal is in line with the argument of Amaral and Meurers (2011: pp. 9–11) that in order to develop effective ICALL systems there has to be a clear identification of the relationship between activity design and restrictions needed to make natural language processing of learner responses tractable and reliable. As the authors propose, despite the most straightforward way to constrain learner production is by explicitly requiring the learner to use certain linguistic constructions, it is more challenging to "investigate how the input can be constrained implicitly in order to provide more space for negotiation of meaning".

In this context, the notion of viable processing ground introduced by Bailey and Meurers (2008: pp. 107–108) is crucial. The processing ground is that set of FL learning activities that (i) combine elements of comprehension and production, (ii) are meaningful and suitable for an ICALL setting, and (iii) require responses that exhibit "controllable" linguistic variation. However, for this to be practical, we need to know the kinds of activities that can be actually found in the viable processing ground. We propose a methodology to classify and analyse the kinds of activity designs that can be correlated with particular expected ranges of responses and particular assessment needs.

Our methodology is conceived as part of a framework to design ICALL materials that is informed by principles of the design of robust NLP resources for the processing of natural language, and particularly for the processing of learner language. Our approach is informed by relevant theoretical frameworks for activity characterisation in the fields of Task-Based Language Teaching and for the characterisation and design of test tasks under TBLT approaches. The relevant concepts in both areas will be introduced in Chapters 3 and 4. With such an interdisciplinary strategy we pursue to generalise the practice and the principles followed by NLP and FLTL researchers in those projects where there was a true cooperative process, where TBLT provided well-defined designs with clear sets of linguistic constructions that "facilitate[d] the restriction to a linguistic domain which is 'manageable'" (Schulze, 2010: p. 79).

2.4.2 An autonomous use of ICALL in class

Our second goal investigates **the autonomous use of ICALL in class**, that is, in foreign language courses in real-world instruction contexts. Feedback provided with NLP technology significantly improves the level and fruitfulness of the student-textbook interactions (Heift and Schulze, 2007: p. 3 and pp. 25–29, Nagata, 2010: p. 461). However, ICALL materials, or ICALL systems, are still used in very few instruction settings. As we described, the implication of the actual teachers in the design of development of these systems is as significant as the collaborative and cross-disciplinary nature of the different areas of expertise used. The development of an authoring tool and the corresponding methodological aids thus is one of the necessary steps to facilitate the integration of ICALL in FL instruction contexts.

This goal is also linked to one important argument in the current research in CALL and ICALL, namely that CALL materials are more easily integrated in the course program when they tend to be designed "with the needs and resources of the individual learner in mind" (Levy and Stockwell, 2006: pp. 11-12). This argument is also claimed by researchers in ICALL, as in Amaral and Meurers (2011: p. 4), who defend that "pedagogical considerations and the influence of activity design choices" condition the successful integration of ICALL systems into FLTL practice.

2.5 Chapter summary

In this chapter, we introduced ICALL as a melting pot at the crossroad of several research disciplines. We saw that ICALL research requires expertise from areas as diverse as Second Language Acquisition, Psycholinguistics, Human-Computer Interaction, User Modelling, Foreign Language Teaching and Learning and Natural Language Processing – the last two being central to this thesis.

We introduced the main issues involved in the development of Intelligent Language Tutoring Systems by providing a detailed description of the seminal work on this topic. We characterised and compared the main features of the three Intelligent Language Tutoring Systems that present a longer and more sustained trajectory both in theory and in practice. We characterised their contexts of use and saw the importance of having the teachers involved from the beginning as a guarantee for the usefulness of the systems. We saw there was substantial progress in the ability to check for the correctness and well-formedness of learner responses, as well as on the adaptation of the system to the learner's performance and its level.

As we are most interested in approaches that integrated ICALL solutions in language instruction settings following communicative approaches to language teaching, we reviewed the research that in one way or another has prioritised the pedagogical goal of the resulting ICALL applications. Particularly, we described the details of ICALL systems that prioritised their technical solutions as much as they prioritised their integration in instruction settings following a Task-Based Language Teaching approach.

We reviewed the first and only study that attempted to characterise ICALL activities both in terms of pedagogical features and in terms of computational complexity in order to determine the range of FL learning activities that are pedagogically meaningful and computationally feasible. This led to the introduction of the concept **viable processing ground**. Making the viable processing ground concrete in form of a pedagogically- and computationally-informed activity characterisation framework is one of the goals of this thesis, namely the one tackled in Part III.

We emphasised the importance of involving teachers in the creation of CALL materials, and we described the little research that was actually carried out in this respect. This grounded an essential part of the second goal of the thesis, namely to empower teachers in real-world instruction settings to produce ICALL activities without the need to be trained in programming. This goal supposes an effort to transfer to teachers in real-world instruction settings the knowledge necessary to understand what is NLP capable of, as well as to understand their professional context and needs. This is the focus of Part IV.

Part II

Background

In the [then] recent broad *Survey of the state of the art in human language technology* (Cole et al. 1996), there is not a single word about (human) language learning. Similarly, CALL contributions to the biennal international conference on computational linguistics (COLING) have been next to nonexistent. [...] Thus, while certainly not part of the core of NLP, CALL seems not to have a place even in its periphery. [...]

The power of CALL (Pennington 1996), which, according to the back cover blurb, [...] "is destined to be the standard reference on CALL and the textbook of choice for teacher training courses covering the use of technology in language learning", contains basically nothing on the uses of NLP in CALL.

Chapelle (1997, 1999, 2001) is not optimistic about the contributions of AI/NLP to CALL, although at least in her 2001 book, the NLP work that she reviews [...] is in most cases more than a decade old, and sometimes more than two decades, in a field which has seen very rapid development in the last ten years.

"What have you done for me lately?
The fickle alignment of NLP and CALL"
NLP in CALL – New Light Penetrates or No Longer Pertinent?
3rd Pre-conference Workshop at EUROCALL 2002 in Jyväskylä (Finland)
Lars Borin (2002: p. 2)

Chapter 3

Natural Language Processing

This chapter presents the methodological and technical background of this thesis from the perspective of Natural Language Processing. We introduce its object of study and its applications to real life. Our presentation introduces three relevant NLP issues for the purpose of this thesis: the technical approach, domain adaptivity, and the robustness of processing strategies. These three concepts impact the strategy followed to develop and use the NLP tools for the processing of learner language. The approach determines the way linguistic information is abstracted: on the basis of human insights and in form of linguistic principles, or on the basis of mathematically-sound algorithms capable of abstracting linguistic properties from annotated data.

Since an unrestricted approach to human language understanding by means of computer-based language processing is today unfeasible (by "unrestricted approach" we mean the possibility of having any given text processed and "understood" by a computer), the notion of domain and domain-specific NLP strategies is introduced. We review how domain adaptivity plays a central role in the usefulness and the feasibility of implementing NLP-enhanced real-life applications.

Finally, robustness is necessary for the stability of system behaviour, since we cannot afford a real-life application to break down even if the analysed language presents non-standard characteristics. We review the different techniques used in NLP to process ill-formed language, the term used in Linguistics and Computational Linguistics to refer to linguistic objects that do not comply with the standard grammar, or the generally accepted conventions, of the language – an abstraction of the linguistic competence of a native-speaker. In particular, we present the NLP research focusing on the analysis of learner language, one of the types of ill-formed language that received attention from the NLP community.

3.1 Fundamental concepts in NLP

According to Jurafsky and Martin (2009: p. 35), processing human language automatically is the principal goal of Computational Linguistics (in Linguistics), Natural Language Processing (in Computer Science), Speech Recognition (in Electrical Engineering) and Computational Psycholinguistics (in Psychology), four research areas that present significant overlap.

For our purposes, speech and language processing can be defined using Jurafsky and Martin's words: "to get computers to perform useful tasks involving human language, tasks like enabling human-machine communication, improving human-human communication, or simply doing useful processing of text" (2009: p. 35). In our case, we focus on written language. More concretely we focus on written learner language.

3.1.1 Approaches to processing natural language

The two basic approaches for processing natural language are usually referred to as **symbolic and stochastic approaches** (Jurafsky and Martin, 2009: p. 44).

Symbolic approaches to NLP are based on algorithms that allow humans, usually computational linguists, to write dictionaries or rules that determine what kind of linguistic operations can be performed on a text. These operations usually imply the linguistic analysis of the words to obtain the corresponding morphological, syntactic, semantic or pragmatic information. According to Jurafsky and Martin (2009: p. 44), symbolic approaches to NLP are related to research lines such as formal language theory, reasoning and type logic.

Stochastic approaches apply data-driven techniques to linguistic tasks using likelihood mathematics and prediction models. These approaches consist in extracting linguistic knowledge from data, usually manually annotated by experts, by means of algorithms that generalise word behaviour on the basis of some sort of distributional property grasped by mathematical principles.

Reasons to choose between one or the other are often related to the size of the data to be handled, as well as the complexity of the phenomenon to be tackled. On the one side, NLP tasks that are well understood in terms of linguistics or that can be reasonably abstracted into lexico-grammatical patterns by NLP specialists are more convenient for symbolic approaches. This is also true for tasks for which the amount of data is low. By contrast, tasks for which there is a large amount of (usually) annotated data, or which are not easily explained in terms of linguistics or lexico-grammatical patterns, are typically tasks suited to stochastic approaches.

Although in the beginning NLP researchers tended to work following either one approach or the other, in the 1990s researchers started to work on hybrid solutions for natural language processing, that is, NLP solutions that used both symbolic approaches and stochastic approaches (see Resnik, 1995, Padró, 1998). The use and implementation of hybrid approaches is still a hot topic today (see for instance the latest EACL workshop on *Innovative hybrid approaches to the processing of textual data*, http://www-limbio.smbh.univ-paris13.fr/membres/hamon/hybrid/).

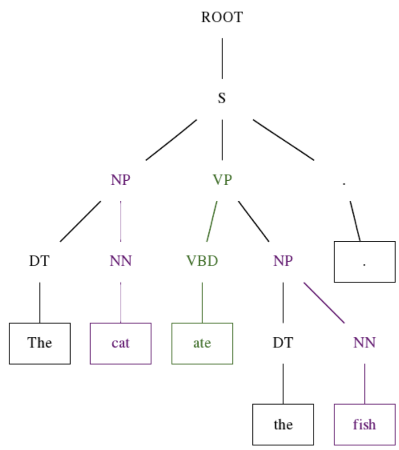
Researchers in the field discussed the advantages and disadvantages of one or the other (see for instance Tapanainen and Voutilainen, 1994, and Voutilainen and Padró, 1997). This thesis does not make a point in this respect: We argue that, if properly designed and implemented, both approaches are equally useful and efficient. However, under certain circumstances – constraintness of the task, availability of large or annotated corpora, cognitive complexity of the task, and so on –, applying one approach or the other might be more efficient. What is relevant to us, in line with Jurafsky and Martin (2009: p. 36), is that "what distinguishes language processing applications from other data processing systems is their use of *knowledge of* *language*", which corresponds to knowledge in phonetics and phonology, morphology, syntax, semantics, pragmatics and dialogue.

3.1.1.1 Deep versus shallow NLP processing

The result of applying NLP techniques to a text is a linguistically analysed text, which is often an intermediate step before an applied task (from machine translation to phone-based banking) can be performed. A typical full syntactic analysis for a sentence such as *The cat ate the fish.* is represented in Figure 3.1.[1] The NLP processing techniques that provide such full-fledged morphosyntactic parses are commonly referred to as deep parsing techniques. Higher levels of analysis might also be tackled, but this does not make a difference in the point discussed here.

```
(ROOT
  (S
    (NP (DT The) (NN cat))
    (VP (VBD ate)
      (NP (DT the) (NN fish)))
    (. .)))
```

(a)



(b)

Figure 3.1: Full syntactic parse of a sentence in parenthetic and tree representation.

Deep parses include all types of relations and interdependencies between the words in the sentences. For instance, in Figure 3.1, the subject *the cat* is described as consisting of a determiner and a noun that, together, form a noun phrase, of which the noun is the head, as represented in the graphical tree in Figure 3.1b. Deep parsing creates too many difficulties for the underlying algorithms, mainly related with the high degree of ambiguity and the complexity of the linguistic structures, though it is also true that certain tasks do require it (e.g. prepositional phrase attachment or coordination ambiguities, Jurafsky and Martin, 2009: p. 451).

In the early 1990s, Abney (Abney, 1991, 1996) introduced the concept of partial, or shallow, parsing as well as the notion of chunk, which he defined as clusters of words that correspond in some way to prosodic patterns (1991: p. 1). Although he applied shallow techniques to syntactic parsing, the term is nowadays generally used any kind NLP task implying linguistic annotation for which a less complex analysis suffices – and replaces a full linguistic analysis.

[1] Parsed with the Stanford Parser at http://nlp.stanford.edu:8080/parser in June 2012. Visualisation through http://www.ark.cs.cmu.edu/parseviz/.

A typical partial or shallow parse of the sentence *The cat ate the fish* is represented in Figure 3.2.[2] This type of analysis ignores the relationships between determiners and nouns, as well as the relations between the verb phrase and the two noun phrases.

[NP The cat] [VP ate] [NP the fish] .

Figure 3.2: Partial syntactic parse of a sentence in parenthetic representation.

Shallow approaches simplify the parsing strategy in that only the amount of information needed to complete the task will be extracted. Moreover, NLP practice has shown that this simplification of the analysis facilitates the application of NLP techniques to real-life tasks for which a complete processing of the text is not needed. This is the case of Information Extraction tasks, where templates requiring specific data must be completed (Jurafsky and Martin, 2000: p. 385–386, Jurafsky and Martin, 2009: ch. 22).[3]

An important advantage of shallow approaches to NLP is the use of cascades of finite-state automata (Abney, 1996, Jurafsky and Martin, 2009: pp. 450–451), which are much more efficient than standard parsing algorithms. Of course, efficiency is improved at the cost of coverage, but the tasks for which it is applied benefit from the trade.

3.1.2 The domain of application

From the very beginning the processing of human language has been applied to tasks whose objective is not the linguistic analysis of language but a more practical communication-oriented purpose. The oldest application of NLP is machine translation, but many other applications have become popular too: language checking, automatic summarisation, term extraction for the creation of glossaries or dictionaries, phone-based services or chatbots allowing for machine-human interaction, question answering, information extraction for database population or template filling, sentiment analysis, and many more.

A crucial aspect in any such application is the domain, in an NLP-specific sense of the word. The domain determines the topics, the linguistic objects (words and structures), and the types of texts that we will find in the language object of the application. For instance, a task of machine translation will pose different challenges if applied to the translation of newspaper text, the translation of technical documentation (Mitamura et al., 1993), or the translation of parliament proceedings (Schwenk, 2007). The domain affects a series of NLP linguistic tasks that are usually performed to tackle the above mentioned non-linguistic tasks, such as morphological analysis, part-of-speech tagging, syntactic parsing, semantic role labelling, semantic analysis, discourse analysis, and so on.

[2] Parsed with the Shallow Parser from the Cognitive Computation Group from the University of Illinois at Urbana Champaign http://cogcomp.cs.illinois.edu/demo/shallowparse/results.php in June 2012.

[3] A typical Information Extraction task is the analysis of, say, economic newspaper news in order to extract a list of the companies that merged, bought or were sold to other companies.

This notion of domain in NLP has a straightforward counterpart in Corpus Linguistics, namely in research that investigates the variation of linguistic phenomena over different genres or registers of language as in Francis and Kučera (1982), Biber (1993), and Roland and Jurafsky (1998) – to mention but a few. The domain usually constrains the possible interpretations of linguistic elements, as well as many other relations, combinations and operations on the basis of their possible interpretations.

As way of example, Daumé III and Marcu (2006: p. 104) explain how in a corpus of financial and economic newspaper text the most frequent reading of the word *monitor* corresponds to its verb reading. In contrast, in a corpus of technical writing the most frequent reading of this word is the noun reading. This has consequences on the way the word is tagged with a part-of-speech, on the kinds of words it co-occurs with, on the kinds of syntactic relations it can build, on the way it is translated to other languages, and so on.

Examples of the importance of domain in NLP tasks can be found in the literature on syntactic parsing (Gildea, 2001; Daumé III and Marcu, 2006; Plank and van Noord, 2010), machine translation (Rosenfeld, 1996; Schroeder, 2007), semantic role labelling (van der Plas et al., 2009; Dahlmeier and Ng, 2010), or sentiment analysis (Pang et al., 2002; Conrad and Schilder, 2007; Binali et al., 2009), to mention a few of the NLP tasks we presented. The importance of domain adaptation is reflected in investigations on how to adapt NLP strategies to domains using semi-supervised and unsupervised techniques to reduce the high costs in time and effort of manual data annotation – see for instance the proceedings of the workshops Domain Adaptation in NLP in 2010 and the NIPS 2011 Domain Adaptation Workshop.[4]

3.1.2.1 The domain in foreign language teaching and learning

In foreign language teaching, a domain corresponds to the language that can be elicited by the activity instructions, the proficiency level of the learners, and probably other sociocultural and pedagogical factors. In a sense, the FL learning domain can be more restrictive than the domain in other NLP tasks. The task or activity instructions delimit the nature of the language elicited, the topic of the conversation or the text, and the text genre that is adequate to the emulated communication setting or that is coherent with the pedagogical goals.

This idea of domain is present already in Weischedel et al.'s seminal work, where the authors state: "Since the instruction involves a text and a particular set of questions, *a complete semantic and contextual model is possible for each text*" (1978: p. 227, my italics). Their ICALL system pursues the assessment of reading comprehension activities, and they see in each of the texts that constitute the input data for the learner a micro-world that can be modelled to assess both meaning and form (1978: p. 233–237). This idea of a "natural" restriction of the domain of application thanks to the constraints imposed by activity design is reflected in the ICALL literature (Quixal et al., 2006, Amaral and Meurers, 2011: p. 11).

[4]Domain Adaptation in NLP 2010 collocated with the Annual Conference of the Association for Computational Linguistics: http://sites.google.com/site/danlp2010/; and the NIPS 2011 Domain Adaptation Workshop, collocated with Annual Conference on Neural Information Processing Systems http://sites.google.com/site/nips2011domainadap/references.

3.1.3 Robust NLP tools

Although NLP is principally oriented to analyse well-formed language structures, it is very common in NLP applications to use strategies to handle deviant linguistic structures – particularly if the application is intended to solve real-world problems. According to Menzel (1995: p. 20), ill-formed input data should have little or no impact on NLP systems if the deviations from the standard language are minor, a behaviour for which he uses the term **robustness**.

Menzel (1995: p. 20) defines robustness as "a kind of monotonic behaviour, which should be guaranteed whenever a system is exposed to some sort of non-standard input data".[5] Robustness is the system's indifference to factors such as uncertainty of real-world input (transcribed or written text), speaker variance (idiolect, dialect, sociolect), erroneous input with respect to standards, insufficient competence (of the processing system), or resource limitation due to parallel execution of several mental activities. While the human brain is capable of ignoring irrelevant information and retaining relevant information under a combination of such conditions, technical solutions "are likely to have serious problems if confronted with only a single type of distortion, apart from the fundamental difficulties to supply the desired monotonic behaviour" (Idem).

Learner language often includes non-existing words, existing words used in an inconsistent or incorrect manner, words incorrectly spelled, sentences with missing parts that should be added to make them comprehensible not just syntactically correct, and so on. Again, humans, teachers or instructors, can usually interpret the intended meaning of a sentence or text even if it contains errors, evaluate whether it complies with the pedagogical goals, provide adequate feedback, and more. However, computers, even if they are programmed to do so, cannot always do it, because not every possible disrupting factor may have been foreseen.

A distinctive aspect of the kind of robustness that is desirable in an NLP system to be included in an ILTS is that it must be able to not only process any given fragment of text, but also has to characterise its linguistic properties to provide pedagogically and linguistically motivated feedback. In other NLP tasks (e.g., machine translation, sentiment analysis, speech recognition) systems must produce an output consistent with the expected functionalities: a translated text, a text correctly classified as a positive or a negative opinion, or a correctly transliterated text. They "only" need to handle ill-formed language, as long as the final result is appropriate.

In analysing the behaviour of syntactic parsers Foster (2007: p. 129) defines four increasingly complex levels of information to which the robustness of a parser can be correlated: some analysis, correct analysis, correct analysis and grammaticality judgement, and finally correct analysis, grammaticality judgement and error correction. Only the two latter levels of information are useful for ICALL, even if not sufficient, since it is not always a matter of grammaticality. An ILTS designed for the instruction of communicative language teaching, not only would require these two levels of robustness at the level of syntax, but also it would require them at the levels of morphology, morphosyntactic features of lexical items, semantics, and

[5]Note here the term input data is used to refer to the input to the NLP system, not to the input for the learner to complete an activity.

pragmatics.

3.2 Analysing learner language

According to Brown (2007: p. 256), learner language is "neither the system of the native language nor the system of the target language, but a system based on the best attempt of learners to bring order and structure to the linguistic stimuli surrounding them". This is often called interlanguage and it includes a notion of an evolution through different developmental stages (2007: pp. 255–256). Learner language is systematic, and its features and characteristics can be described as it can be done with any other language. However, the development of NLP tools is traditionally based on the characterisation of language as spoken or written by native speakers.

The task of developing NLP tools specifically for learners is not feasible to date, since the current studies that describe the different developmental stages of the acquired language are not thorough enough, nor sufficiently big annotated corpora exist. Even if developing stage-wise learner-specific NLP tools were possible, the strategies to ensure the selection of the appropriate resources for the analysis of a particular response of a particular learner to a particular activity would still have to be developed. This is an overwhelming task that research in ICALL solves by implementing strategies to handle a range of deviations with respect to the normative standard of a language.

3.2.1 Symbolic approaches to process ill-formed language

There are basically two approaches to the development of NLP tools for the processing of ill-formed language: the mal-rule approach, and the constrain relaxation approach. Both are types of symbolic approaches to language processing. A third possible approach to the analysis of learner language is any stochastic approach to NLP, but, as we explain later, the robustness that stochastic systems provide is not always interpretable linguistically speaking.

3.2.1.1 Mal-rule approach

The main characteristic of the **mal-rule approach** is that it explicitly models ungrammatical structures in form of grammar rules, so that ill-formed language can be successfully parsed. To exemplify it, let us take the rule in (3.1). This is a standard rule for the analysis of noun phrases consisting of a determiner, an adjective and a noun. The rule includes a variable, V1, that requires that the three words agree in gender and number for it to be successfully applied. Such a rule would parse a sequence like *these big peaches*$_{NP}$, but not a sequence as **this big peaches.*

$$NP \Rightarrow Det_{Agr::V1} + Adj_{Agr::V1} + Noun_{Agr::V1} \tag{3.1}$$

To parse a sequence like **this big peaches*, a mal-rule like the one in (3.2) could be used. The rule states that a sequence consisting of a determiner, an adjective and a noun in which the determiner is singular and the adjective and the noun are plural. Such a rule would successfully parse a sequence such as *these big peaches*$_{NP_{NumWrong}}$.

$$NP_{NumWrong} \Rightarrow Det_{\substack{Gen::Any \\ Num::Sg}} + Adj_{\substack{Gen::Any \\ Num::Plu}} + Noun_{\substack{Gen::Any \\ Num::Plu}} \tag{3.2}$$

The mal-rule approach has different implementations. Dini and Malnati (1993) identify three different versions of the mal-rule approach: the rule-based approach, the meta-rule based approach, and the preference-based approach. The rule-based approach consists of two different grammars, one containing the standard rules for the processing of ill-formed language and another containing the rules for the processing of ill-formed language. The meta-rule approach captures the explicit modelling of ill-formed sequences in a separate set of rules that determines which rules and which linguistic features in the standard grammar can be violated and which not, hence the name *meta*-rules. The preference-based rule consists of two modules containing syntactic phrase re-write rules and syntactic structure building rules. While the re-write rules overgenerate because they do not restrict many morphosyntactic features, tree building rules record the morphosyntactic features that are violated or not according to a list of pre-established features. The system monitors and stores the modified features each time a modification is required.

3.2.1.2 Constraint relaxation approach

As for the **constraint relaxation approach**, it is based on a modification of the parsing algorithm so that certain conditions in the grammar rules can be relaxed. This allows for the construction of a less restrictive grammar on the basis of a standard grammar, and provides a principled connection between constructions accepted by either grammar (Douglas and Dale, 1992: p. 469). For explanatory reasons, we will present and use in this section an example of the relaxation approach to parsing presented by Douglas and Dale (1992). Further references on relaxation approaches can be found in Weischedel and Black (1980), Kwasny and Sondheimer (1981), or Richardson and Braden-Harder (1988).

In (3.3) we present a rule that parses a noun phrase consisting of a determiner and a noun. It is similar to the rules in (3.1) and (3.2), without the adjective. The rule has six numbered constraints: Constraint [1] gives the resulting element a label, Noun Phrase; Constraints [2] and [3] determine respectively the category of the first and the second element; Constraint [4] controls that the determiner ends with the letter *n* only if the noun starts with a vowel or not; Constraint [5] requires agreement in number between the determiner and the noun; and Constraint [6] assigns the resulting phrase the number of its head, the noun.

$$
\boxed{
\begin{array}{ll}
X0 \;\Rightarrow\; X1\,X2 & \\
\quad \langle X0\ cat\rangle \;= \text{NP} & [1] \\
\quad \langle X1\ cat\rangle \;= \text{Det} & [2] \\
\quad \langle X2\ cat\rangle \;= \text{Noun} & [3] \\
\quad \langle X1\ agr\ en\rangle = \langle X1\ agr\ vow\rangle & [4] \\
\quad \langle X1\ agr\ num\rangle = \langle X1\ agr\ num\rangle & [5] \\
\quad \langle X0\ agr\ num\rangle = \langle X2\ agr\ num\rangle & [6] \\
\text{Relaxation level 0:} & \\
\quad \text{necessary constraints: } \{2,3,5,4,1,6\} & \\
\quad \text{relaxation packages: } \{\} & \\
\text{Relaxation level 1:} & \\
\quad \text{necessary constraints: } \{2,3,1\} & \\
\quad \text{relaxation packages:} & \\
\quad\quad \text{(a) } \{5,6\} \text{ : Det-noun number disagreement} & \\
\quad\quad \text{(b) } \{4\} \text{ : a/an error} &
\end{array}
}
\tag{3.3}
$$

As shown in (3.2.1.2), different levels of relaxation are indicated in a separate area. For level 0, no relaxation is allowed, while for level 1, three of the constraints have been moved to the area where relaxation packages (lists of relaxation statements) are stated. Constraints [5] and [6] are grouped into one package. This is a designer decision to ensure that no number is assigned to the NP if there is disagreement between determiner and noun in this feature. As for Constraint [4], it is alone, but it is ranked below the other two, implying that it will not be checked if Constraints [5] and [6] fail to be applied. The main difference with respect to the mal-rule approach is that the constraint relaxation approach requires no specific rules to develop the system.

3.2.1.3 Pros and cons

According to Heift and Schulze, with any of the mal-rule techniques it is difficult to handle multiple errors because it is a very localised approach to processing (2007: p. 40). Moreover, the authors think mal-rule approaches have the disadvantage of presenting redundant linguistic information, and require an extensive anticipation of whatever ill-formed structures are to be parsed. However, in our opinion, both techniques require a considerable effort in anticipating the kinds of deviations (or errors) to be properly handled. Independently of the approach followed, the particular linguistic features to be analysed and used in either explicit rules or in relaxation statements must be carefully selected and applied.

In terms of number of rules, a pure mal-rule approach certainly requires the writing of more rules – far more than doubling or tripling the original, depending on the deviations to be implemented and the expressive power of the formalism. In contrast, as described in Heift (2003), in implementing a relaxation approach for

E-Tutor, the rules did not have to be modified, but a number of components in the architecture had to. This was the case for the lexicon and the configuration of general unification principles (in an HPSG-like implementation).

As for their application in ICALL, both approaches are used in the implementation of the three ICALL systems that we analysed in detail in Section 2.1.4. *ROBO-SENSEI* (Nagata, 2002, 2009) and *TAGARELA* (Amaral, 2007; Ziai, 2009) use a mal-rule approach in some of their modules, while E-Tutor (Heift and Nicholson, 2001; Heift, 2003) uses both the mal-rule and the relaxation approach.

3.2.2 Stochastic approaches to detect deviations from the norm

Stochastic approaches to NLP provide the robustness inherent to the generalisation capability of the underlying mathematical principles. However, as stated in Foster (2007: p. 130), the robustness provided by stochastic syntactic parsers is not informative enough for ICALL systems. In the author's words, "treebank-trained statistical parsers are generally agnostic to the concept of grammaticality", and the fact that they are trained on edited text, such as *The Wall Street Journal*, can be interpreted as a weakness for the processing of ill-formed input. See, for instance, Wagner et al. (2009) for a discussion on the limits of using and combining rule-based and/or stochastic syntactic parsers to judge sentence grammaticality.

However, in the last decade NLP researchers investigated the use of stochastic approaches for the detection of errors. Such approaches do not target the analysis of ill-formed text, but instead aim to classify text sequences (phrases, word combinations, usage of prepositions or determiners) as compatible with standard native-speaker language models or incompatible and, therefore, deviant or at least rare. Interestingly, many of these techniques are not being implemented for the detection of errors at the syntax level, but rather at the level language use.

The seminal work in this area was carried out by Golding and Schabes (1996), Golding and Roth (1999) and Chodorow and Leacock (2000). These researchers proposed supervised and weakly supervised error detection strategies to detect confusables (e.g., *affect* vs. *effect*) and word usage (e.g., *concentrate on* as opposed to **concentrate in*) requiring a context-sensitive error detection strategy. The authors describe systems that exploit statistics based on the distribution of the target words in general or word-specific corpora. In slightly different implementations, such strategies allow for the detection of word usages in contexts where their occurrence probability is low – an susceptible to being an error or deviation from standard language.

Similar and improved approaches were developed since then for the detection of errors on the basis of POS-tag distribution (Bigert and Knutsson, 2002), and for the detection of determiner and preposition usage, (Chodorow et al., 2007; Tetreault and Chodorow, 2008; Gamon et al., 2008; Felice and Pulman, 2009; Elghafari et al., 2010). None of these systems has been integrated in ICALL systems, though some were applied to so-called writer aids (Gamon et al., 2008; Bigert and Knutsson, 2002) or holistic essay scoring software (Burstein et al., 2003). However, statistical techniques are data-intensive, and no such techniques can be applied if the training or evaluation data cannot be obtained (automatically or manually) from annotated large corpora. In the absence of annotated data, they must be collected, created them from scratch, or artificially generated (Foster, 2007, Wagner et al., 2009).

3.3 Chapter summary

We introduced the object of study of Natural Language Processing. We showed how language analysis can be tackled using symbolic approaches based on hand-crafted written linguistic grammars, or stochastic approaches that exploit statistical observations about large amounts of (annotated) data. Both approaches have proven useful and effective in real-life applications under particular circumstances, and both can be combined to develop so-called hybrid approaches.

Since we are interested in analysing learner language, whose properties are different from the standard type of language NLP tools are typically developed for, we presented shallow NLP techniques. Such techniques allow for a simplified linguistic analysis of texts without compromising the result for certain real-life tasks that require automatic analysis of language, and are more stable computationally speaking. Shallow approaches to language processing are employed in the ICALL systems used in a sustained manner over the past decades – see Chapter 2.

We introduced the notion of domain in applied NLP research, since the domain characterises the topics and the language that a particular application faces. Since task-based language teaching and learning aims to emulate communicative settings in real life as a means to put the learner in the need to use particular linguistic structures, the notion of domain in NLP naturally correlates with the notion of target language use setting in task-based instruction – a notion presented in the following chapter.

We explained the different approaches to the analysis of ill-formed language. Since most NLP tools are developed on the basis of native speaker language rules, symbolic approaches to NLP were complemented with strategies to overcome the inherent difficulties of parsing language that deviates from the norm: the mal-rule approach and the constrain-relaxation approach. Both approaches are used in existing ICALL systems. Robustness is the characteristic that allows NLP resources to cope with learner language, which is prone to containing lexical and grammatical elements deviating from native speaker language.

This chapter described the methodological and technological background of the research framework presented in Chapter 6, which provides the pedagogical and computational characteristics of our research context.

Chapter 4

Foreign Language Teaching and Learning

This chapter introduces the theoretical and methodological background of the thesis from the perspective of Foreign Language Teaching and Learning. We start with an introduction of the pedagogical approach in which we frame our research, namely Communicative Language Teaching and in particular Task-Based Language Teaching. In this context, we pay special attention to Form-Focused Instruction, an approach that emphasises the need to draw the learner's attention to formal aspects of language while fostering communicative competence in a foreign language.

After this introduction, we review research in three different areas: (i) the concept "task" and the way it has been applied to the design of task-based instruction materials, (ii) the assessment of learner production, focusing on types of feedback and their effectiveness, and (iii) the effects of feedback on FL learners using CALL materials.

For the concept of task, we review the studies that define task and classify tasks in terms of pedagogical criteria, such as the extent to which tasks enhance the learner's communicative competence, the freedom of the learner in the selection of the linguistic resources to respond to the task demands, or the similarity of a given task with real life communicative outcomes and/or linguistic or cognitive processes. To link the theoretical concept of task with FLTL practice, we present a framework for the development of task-based instruction materials. The characterisation of tasks and the alignment with task-based approaches to material design are the pedagogical basis for our framework to develop ICALL materials in the following chapters.

To establish evaluation needs and requirements for NLP-based assessment tools, task assessment criteria must be defined. For this purpose, we review the FLTL literature on assessment types and the characterisation of assessment materials. We describe summative and formative assessment, and we present a framework for the characterisation of language tests that qualify as tasks. Such a framework allows for a detailed characterisation of the properties of the language and the communication strategies to show up in a given task. This set of linguistic and communicative properties of tasks will link to the notion of NLP domain presented in the previous chapter.

4.1 Modern instruction of foreign languages

According to Brown (2007: p. 18), modern approaches to language instruction are tightly correlated with theoretical disciplines such as psychology, psycholinguistics and second language acquisition.[1] In his view, Communicative Language Teaching is the approach that best captures the different "pedagogical springs and rivers of the last few decades" by offering an eclectic blend of the contributions made by previous approaches to language teaching.

Before the 20^{th} century languages were taught following the Classical Method, later known also as the Grammar Translation Method. This method, which was inspired by the way Latin and Greek had traditionally been taught, consisted in learning grammatical rules, memorising vocabulary, declensions and conjugations, translating texts and doing written exercises. In one way or another this method has been used for many centuries worldwide and is still used today (Brown, 2007: p. 14–20) .

In the 20^{th} century, several methods appeared: e.g., Direct Method, Series Method, and Audiolingual Method (Brown, 2007: p. 17). However, current language teaching practices can be characterised as absent of "proclaimed 'orthodoxies' and 'best' methods" (2007: p. 18, original quotes). Current practices argue against the use of "instant recipes" and encourage teachers to develop a "principled basis upon which [they] can choose particular designs and techniques for teaching a foreign language in a specific context" (Ibid.).

CLT should be understood as an approach rather than as a method (Brown, 2007: p. 18, Richards and Rodgers, 2001). Today, in the Western pedagogics tradition, it is the most common approach in FL instruction (Brown, 2007: pp. 241). Brown sees CLT as a "unified but broadly based theoretical position about the nature of language and of language learning and teaching" – see also Breen and Candlin (1980), Savignon (1983), Nunan (2004), and Ellis (2005) on the conceptualisation of CLT. According to Brown (2007: p. 241), CLT is fundamentally characterised by the following:

1. Classroom goals focus on all components of communicative competence, not only grammatical or linguistic competence.

2. Language techniques are designed to engage learners in the pragmatic, authentic, functional use of language for meaningful purposes. Organisational forms are not central, but enable the learner to accomplish those purposes.

3. Fluency and accuracy are seen as complementary principles underlying communicative techniques, although at times fluency may be prioritised to keep learners meaningfully engaged in language use.

[1] Brown (2007) is not the standard reference in Second Language Acquisition research to refer to the foundations of CLT and TBLT. We take this book as a reference because its target is the actual teacher or teacher trainees, and teachers are the focus of our research in Part IV of the thesis. In our opinion, the book provides a detailed and thorough critical interpretation of many of the findings in SLA and FLTL research, and many other fields related to FLTL practice.

4. In the classroom, learners must use the language productively and receptively in unrehearsed contexts.

As an approach with its roots in the notional-functional syllabus, Brown says that CLT subsumes grammatical structure under language functional categories such as reporting, asking permission, introducing oneself and other people (2007: pp. 241–242). The emphasis that CLT places on functional categories results in a need to set up instruction environments in which learning activities focus on meaning, that is, the use of language in context, as opposed to focus on form, that is, the systemic properties of a language. This requires authentic language, particularly to improve fluency, but not at the expense of clear, unambiguous, direct communication.

In recent years, a branch of research in SLA has argued for so-called **form-focused instruction** (FFI), an argument increasingly acknowledged in CLT (Brown, 2007: p. 276). According to Spada (1997: p. 73), FFI can be defined as "any pedagogical effort which is used to draw the learners' attention to language form either implicitly or explicitly", an effort that can be planned or spontaneous (Brown, 2007: p. 277). Modern uses of FFI foresee an occasional and carefully selected focus on the formal aspects of language. However, FFI is not the same as focus on forms (note the "s"), that corresponds to the traditional grammar-based way of teaching languages. FFI has been proven as effective or even more effective than the traditional approach to language learning (Long and Doughty, 2011: p. 381).

4.2 TBLT: principles and practice

In this section we present the concept of Task-Based Language Teaching, the concept of task and the criteria provided by FLTL research to determine when a FL learning activity qualifies as a task. Later, we present a practical framework for developing task-based course materials.

4.2.1 Task-Based Language Instruction

According to Brown (2007: p. 242), Task-Based Instruction (TBI) "has emerged as a major focal point in language teaching practice worldwide". TBI is a particular method of syllabus design in which units of work are conceived according to the goals and the needs of a specific task that learners ought to be able to do in their everyday lives.

The definition of task is not free of controversy. Ellis (2003: Ch. 1), Nunan (2004: Ch. 1), and Brown (2007: pp. 242–243) review and offer several definitions of task. We can distinguish a broader and a narrower definition of task. Long (1985: p. 89) is one giving a broader definition of task: "the hundred and one things people *do* in everyday life, at work, and play or in between." Nunan (2004: p. 10) defines task more narrowly: "a piece of classroom work that involves learners in comprehending, manipulating, producing or interacting in the target language while their attention is focused on mobilizing their grammatical knowledge in order to express meaning, and in which the intention is to convey meaning rather than to manipulate form".

In line with Ellis (2003) and Brown (2007), we adopt the narrower definition of task, also known as the pedagogical conception of task. According to this view, tasks occur in classroom, or as part of out-of-class work. Pedagogical tasks can be opposed to target language use tasks, since the latter are understood as the use of language in the world beyond the classroom.

Littlewood (2004: p. 322) proposes a continuum to characterise tasks. This continuum ranges from language teaching activities with a focus on form to language teaching activities with a focus on meaning. We present them incrementally from more form-oriented to more communicative-oriented:

1. **Non-communicative learning** is pure focus on forms (with "s"), that is, on the structures of language, and it might imply substitution exercises, discovery and awareness raising activities.

2. **Pre-communicative language practice** includes language practice with some attention to meaning but not communicating new messages, such as question-and-answer practice.

3. **Communicative language practice** is in the middle of this continuum. It corresponds to practising pre-taught language in contexts where it communicates new information, such as information-gap activities or personalised questions.

4. **Structured communication** activities are those in which language is used to communicate in situations that elicit pre-learnt language with some unpredictability, such as role-play activities and simple problem solving.

5. **Authentic communication** includes activities that imply using language to communicate in situations where the meanings are unpredictable such as creative role-play activities, or more complex problem-solving and discussion.

Littlewood defines the less "communicative" extreme in his continuum as *exercises* or *enabling tasks*, terms respectively used by Ellis (2003) and Estaire and Zanón (1994). At the other extreme of the continuum, he defines *tasks* or *communicative tasks*, respectively used by Ellis (2003) and Estaire and Zanón (1994). According to Littlewood, the middle categories possess properties of both.

4.2.1.1 Analysing the properties of language learning activities

Ellis (2003: pp. 9–10) devoted part of his research to carefully analyse the characteristics that make tasks *communicative tasks*. This research can be used to further clarify pedagogical and linguistic features of tasks. Ellis proposes seven criterial features to identify tasks:

(1) A task is a work plan in the form of teaching materials or ad hoc materials for activities that arise during the course of teaching.

(2) A task involves a primary focus on meaning. That is, it uses language pragmatically. A task incorporates information, opinion or reasoning gaps, that motivate learners to use language to close it.

(3) A task should allow learners to choose the linguistic and non-linguistic resources needed to complete it, and achieve the outcome of the activity. A task creates a semantic space that constrains the linguistic forms learners will use.

(4) A task involves real-world processes of language use to engage in activities emulating the real world, such as completing a form, asking and answering questions, or dealing with misunderstandings.

(5) A task can involve any of the four language skills. It may require learners to read or listen to a text and later on demonstrate their understanding. It may require them to produce an oral or written text. Or it may require them to employ a combination of skills.

(6) A task engages cognitive processes, that is, the workplan requires learners to employ cognitive processes such as selecting, classifying, ordering, reasoning, and evaluating information.

(7) A task has a clearly defined communicative outcome. The stated outcome of the task is a means to determine when participants have completed it.

Ellis (2003: pp. 9–16) uses these seven "criterial features" to classify language learning activities as tasks, focusing on meaning, or as exercises, focusing on form. In his view, tasks are generally preferable to exercises, but there are theoretical grounds to include exercises alongside tasks.

According to Ellis (2003: p. 9), some of these criterial features are common to any instruction material, but a few of them are inherently communicative tasks. In his view, three of them are applicable to all sorts of teaching materials, namely (1), to have a workplan; (5), to involve one or more of the language skills; and (6), to engage cognitive processes. A task should thus present the following features: (2), the focus on meaning; (3), the possibility that learners choose the resources needed to complete the task; (4), involving real world processes; and (7), establishing a clearly defined communicative outcome. However, he stresses that the "key criterion" of *taskness* is no. (2), to have a focus on meaning.

4.2.2 The design of a TBLT syllabus

This section introduces Estaire and Zanón (1994)'s framework for the design of syllabi[2] that employs a task-based instruction method. This framework is a practical guide to the implementation of the TBLT approach that has received wide attention from FLTL researchers and practitioners (see Nunan and Lamb, 1996: p. 47, Ellis, 2003: p. 15, Nunan, 2004, Littlewood, 2004). According to Nunan and Lamb, this approach is appropriate for two particular reasons: it integrates content, objectives,

[2]The word *syllabus* is the British word for the American English *curriculum* (Ellis, 2003).

methodology and evaluation; and it proposes that teachers design tasks and then backward from the target tasks to identify contents, procedures, and instruments for assessment. Ellis (2003: p. 33) values this approach positively because it addresses the topic of fitting tasks into the cycle of teaching. There are other significant references as to the implementation of TBLT in the classroom (e.g., Willis, 1996).

Estaire and Zanón (1994: p. 4) propose a ten-stage development cycle for the generation of task-based units of work:

PLANNING

1. Determine theme or interest area.
2. Plan final task or series of tasks.
3. Determine unit objectives.
4. Specify contents that are necessary or desirable to carry out final task(s).
5. Plan the process: determine communication and enabling tasks that will lead to final task(s); select/adapt/produce appropriate materials for them; structure the tasks and sequence them to fit into class hours.
6. Plan instruments and procedure for evaluation of process and product (built in as part of the learning process).

IMPLEMENTATION

7. Do the unit in the classroom.

A POSTERIORI ANALYSIS

8. Analyse and reflect on the unit in action.
9. Retrospective syllabus: take note and record what actually happened.
10. Plans for the future: modifications to the unit, ideas to recycle content for other units, and ideas for improving effectiveness of learning in future work.

Estaire and Zanón (1994: p. 49) group these ten stages into three more general phases: the planning, the implementation and the a posteriori analysis. The planning phase includes stages 1 through 6 that take place before the unit is actually carried out, and therefore help anticipate the actions and language that will be required for learners to complete the tasks.

The implementation phase includes stage 7, basically the actual use of the materials in the classroom. In a CALL context the use of the materials does not necessarily need to be done in the classroom. The a posteriori analysis includes stages 8 through 10 in which teachers, and learners if desired or appropriate, evaluate, reflect, re-think and re-make the syllabus or the materials.

According to Estaire and Zanón, stages 1 through 3 “lay the foundations for the unit through a general statement of what [teachers] intend to do” (1994: p. 28).

Through them teachers determine the target language use task(s) that the pedagogical tasks must emulate and the objectives pursued.

In stages 4 through 6, unit objectives are materialised in the form of classroom work. In stage 4, thematic contents are narrowed down until we can select the concrete activities that learners are going to do, and particularly the final task. Whatever learners do, and are expected to do, is the main source to determine the linguistic content necessary for the unit. For instance, if learners are supposed to work on creating a newspaper, they must decide whether to include sections on music, sports, or the weather. Each of these decisions influences the linguistic contents, particularly at the lexical level. In determining the linguistic contents, one must think of the language, that is, functions, grammar, vocabulary, discourse features, phonological aspects, and so on, that learners should learn or develop.

Stage 5 requires the designer to organise the "ingredients" in stage 4 and plan the process that will lead learners from the start to the end of the unit of work. Estaire and Zanón recommend here four necessary steps. First, one must decide on the preparatory tasks that help learners prepare for the final task. Within preparatory tasks, they differentiate between enabling tasks and communicative tasks, which we already presented above using Littlewood (2004)'s continuum. Second, one must select, adapt or produce appropriate classroom materials. Third, all tasks will be structured so that the unit's purpose is reached, and this requires a clear procedure and an outcome. And fourth, one must sequence tasks so that they fit into class hours. This must be done coherently and combine communication tasks and enabling tasks with a specific focus on the linguistic system.

Stage 6 is the planning of evaluation instruments and procedures. Estaire and Zanón recommend planning them before the unit is started. The goal of evaluation is to give teachers and learners feedback to determine adjustments and replanning of the work to ensure that learning takes place efficiently and effectively. The authors offer a range of aspects that can be evaluated. As for teachers, they must decide who evaluates what and how this is done. With respect to what is evaluated they mention the process (materials, learning strategies, interaction and participation, etc.), the product (performance, achievement of objectives, etc.), the teacher and the learners. With respect to who evaluates they consider the teacher, the learners and others (such as teachers or peer learners, native-speakers, learners in other courses, etc.). And finally, with respect to how, they propose to use questionnaires, self-assessment, observation, self-observation, tests, and tasks.

Stage 7 is the time to observe and analyse the unit in action, possibly including the learners as observers (e.g., through videos). Stage 8 is an a posteriori analysis of all the information gathered in the previous stages to inform the work of stages 9 and 10. Stage 9, the retrospective syllabus, consists in modifying aspects such as contents, which might have been covered differently as stated in stage 4, or objectives that may have been achieved differently as stated in stage 3. The changes may include modifications to the tasks designed or the materials, or even the evaluation procedures.

Finally, stage 10 is the time to change the unit for future uses, or to select some materials to be recycled. It can consist in taking notes on how things went and what could be observed in learners and in teachers. They propose to pay special attention to the effectiveness of the learning process during this stage.

4.2.2.1 The linguistic contents of tasks

Following the convention in TBLT, Estaire and Zanón suggest specifying the linguistic contents during the development of the units of work (1994: pp. 30, 58, and 63). For each task, they propose filling in a grid including the functional contents, the exponents of functions, the grammatical contents, the lexical contents, and any other contents that are relevant for a particular task, such as phonological or discourse contents, or aspects influencing communicative competence.

In Figure 4.1, we reproduce the table they propose for a task whose thematic aspects are daily routine, free time activities, and personal information. In column 2, there are different exponents of the linguistic-communicative functions used in the task: *I get up at ... (every day)*, *I work from ... to ...*, *How often do you ...*, or *Do you like ...?*

Column 3 shows the grammatical contents learners are expected to use: verb forms in Simple Present included in affirmative, interrogative, or negative sentences, or short answers. This column also requires us to pay attention to the position of frequency verbs and to time expressions. In column 4, they refer to the vocabulary that is expected in the task: verbs referring to daily routine and free time, frequency adverbs, days of the week, parts of the day, and so on.

In developing ICALL materials, this specification table informs us about the kind of language expected. The table details not only the linguistic items and structures that we expect in learner production, but also those aspects that will be most relevant in the assessment of learner production.

1 Functional content Express, ask for and understand information referring to:	2 Exponents of functions	3 Grammatical content	4 Lexical content
Daily routine and free time activities	I get up at… (every day) I work from … to … In the [morning / afternoon / evening] I (usually) On Monday(s) / At the weekend – I .... What time do you .... ? How often do you .... ? How many hours do you .... ? What do you do .... ? Do you (eg work) .... ? Do you like .... ? Yes, I do. No, I don't.	Simple Present (all persons) ◆ affirmative ◆ interrogative ◆ negative ◆ short answers Position of frequency adverbs Time expressions: ◆ *at* (time) ◆ *from … to* ◆ *in the* (parts of day) ◆ *on* (days of the week) ◆ *every day* ◆ *first, then, after that*	Verbs related to daily routine and free time (eg get up, have lunch, work, study, read, play, watch TV, go to bed …) Frequency adverbs Days of the week Parts of the day How often…?

Figure 4.1: Linguistic content specified for a task according to the approach proposed by Estaire and Zanón (1994: p. 63).

4.3 Assessment of learner production

Despite the controversy over the efficacy of oral or written output as part of the language acquisition process (Brown, 2007: pp. 293, and 297–299 and Ellis, 2003: p. 110–115), recent work showed that learner production helps learners gain consciousness of their command of the language being acquired, which enables them to build up a coherent set of knowledge (Swain, 2005, 2000; Swain and Lapkin, 1995; de Bot, 1996). On the basis of this research, Brown argues that learner production in the target language can help learners realise "erroneous attempts to convey meaning" and, through that, recognise their linguistic shortcomings (2007: p. 298).

Brown (2007: pp. 255–257) asserts that language learning is "a process of the creative construction of a system in which learners are consciously testing hypotheses about the target language". This inherently implies that the learner makes mistakes. However, the difficulty of learning a foreign language can be overcome by using a "concerted strategic approach". This strategic approach includes assessment and a "trial and *error*" strategy (Brown, 2007: p. 273–275, original italics). Learner production is corrected and evaluated in instruction contexts, and research shows that learners expect and wish to receive feedback (Brown, 2007: p. 274–276, Chandler, 2003: p. 270), despite the controversy over the efficacy of feedback (Chandler, 2003; Truscott, 2004; Bitchener et al., 2005).[3]

A reasonable position in this respect is found in (Brown, 2007: p. 273):

> Historically, error treatment in language classrooms has been a hot topic. [First] errors were viewed as phenomena to be avoided by overlearning, memorizing, and "getting it right" from the start. Then, some methods [...] took a *laissez-faire* approach to error [...]. CLT approaches, including task-based instruction, now tend to advocate an optimal balance between attention to form (and errors) and attention to meaning.

Following Brown, we assume that error correction is needed, as long as it is adequate to learner style and level. In this sense appropriate types of feedback should include positive, neutral or negative feedback, and affective and cognitive feedback. Feedback is oriented to help the learner gain some knowledge, which is presumably incomplete (more on this topic in Section 4.4).

The assessment of learner production can be achieved through **summative assessment** or **formative assessment**. According to Ellis (2003: p. 312), assessment in TBI must include both types of assessment. While formative assessment is expected to help learners progress in their acquisition of knowledge, summative assessment is expected to show them and the teacher or any other stakeholder how good they are with respect to certain communicative and linguistic abilities at a given point in time.

[3] The research by Chandler (2003), Truscott (2004), and Bitchener et al. (2005) is carried out in the context of composition writing instruction, but is still relevant.

4.3.1 Summative assessment

Summative assessment must be associated with language tests (Ellis, 2003: Ch. 9, Bachman and Palmer, 1996: Ch. 2, Bachman, 1990). Ellis (2003: p. 283–286) proposes to distinguish between two types of tests. System-referenced tests aim to inform about the learner's language proficiency in general, while performance-referenced tests seek to inform about the learner's ability to use the language in a specific context. Ellis distinguishes between direct and indirect assessment. The former involves "the holistic measurement of language abilities involving some kind of task", whereas the latter involves "measuring language proficiency analytically by means of tests of discrete points of language or of specific tests in a task" (*Ibid.*).

Since tasks per se do not provide a measure of the learner's language ability, learner performance must be measured in some way. For this, Ellis describes three possible methods: The first is direct assessment of task outcomes, which is possible in closed tasks that result in a solution that is either right or wrong. For instance, if, as a result of a task, learners must grasp one particular object on a table that has more than one object on it, the result can be directly assessed. If the outcome is a communicative one (a piece of language), it might be more open, but if the response must convey some sort of message, this message must be included in it somehow.

Second, he suggests discourse analytic methods that are based on counts of specific linguistic features occurring in the discourse that result from performing the task. Such methods will relate to the learner's linguistic competence measured in terms of complexity, accuracy and fluency measures. They can also be related to sociolinguistic competence: appropriate use of requesting strategies; or related to discourse competence (e.g., use of cohesive discourse markers).

Finally, there is the external ratings method, which involves an assessor observing the task and making a judgement. This method differs from direct assessment in that the judgement is more subjective, although efforts must be made to warrant reliability. Such a method requires assessment guidelines and possibly a checklist of competencies.

Assessment is an important and very complex process in language teaching and learning, and the measurement of test-task performances has to rely on methods that are valid and reliable (Ellis, 2003: p. 283–286). Because of this Ellis requires complex, qualitative and multidimensional assessments. However, he admits that assessment must be practical and cost-effective, for which assessment procedures must designed according to working conditions and professional expertise.

4.3.1.1 A framework for the characterisation of test tasks

Bachman and Palmer (1996: Ch. 3) present a framework for the characterisation of tasks in tests, whose aim is threefold: (i) to describe the target language use domain (that, is the communicative setting in real life) that will be the basis for the pedagogical task; (ii) to describe test tasks as a means to ensure their comparability and assess their reliability; and (iii) to assess the authenticity of language test tasks (1996: p. 47).

In practical terms Bachman and Palmer's framework should enable teachers to develop test tasks in a principled and objective manner. On one hand, it allows for the comparison of differences and similarities between the test tasks and the corresponding target language use settings in the real-world. On the other hand, it can be used to devise new test tasks that differ from the existing catalogue in a complementary and pedagogically driven manner.

Bachman and Palmer's task characteristics framework

Bachman and Palmer analyse test task characteristics in terms of five different features: setting, test rubric, input, expected response, and relationship between input and response.

As for the **setting**, it comprises physical characteristics such as temperature, seating conditions, lighting, and so on, as well as the participants in it, in addition to the testee, and the time of the day in which the task is to be completed. The **rubric** includes the structure of the test, (items/parts, salience of parts, sequence of parts), the characteristics of the instructions (e.g., language and channel in which they are given), the duration of the test and its items or parts, and, finally, the scoring method, which includes criteria for correctness and a procedure for scoring the response.

As for the **input**, Bachman and Palmer distinguish between format and language characteristics. The input is whatever information the learner is required to process in order to complete the task. The input's format includes the channel (aural, visual or both), or the language (in the native or foreign language). The input's language is analysed in terms of language knowledge and topical knowledge. Language knowledge relates to linguistic aspects such as vocabulary, morphology, syntax, pragmatics (cohesion, rhetorical structure), dialect, register, or cultural references. Topical knowledge relates to the type of information that is part of the input: personal, cultural, academic, technical, and so on.

The characteristics of the expected **response** (as opposed to the actual response) are also analysed by format and language characteristics, exactly as the input is analysed. Moreover, Bachman and Palmer define three types of responses: selected responses (no language product required) and limited or extended production responses. Limited production responses consists of a single word, a phrase, or at most a full sentence. Extended production responses consists of a text that extends somwhere between two utterances and a full text.

The last aspect Bachman and Palmer (1996) propose is the **relationship between input and response**. According to them, this relationship can be measured in terms of reactivity, scope and directness. Reactivity relates to the extent to which input or response affect subsequent input and responses. In this sense, the relationship can be reciprocal, where there is immediate feedback, in the widest sense of the word, that favours interaction between the learner and the interlocutor, or non-reciprocal, where there is no feedback or interaction until the task is finished and evaluated.

As for the scope of the relationship, the authors relate it to the amount of input that must be processed for learners to respond as expected. The scope can be broad, where much input must be processed, or narrow, where the amount of input to be processed is minimal. Finally, the directness of the relationship is related to the degree to which the expected response can be based on information found in or inferable from the input, or whether the learner must rely on information in the context or in his or her own topical knowledge.

4.3.2 Formative assessment

According to Ellis (2003: p. 312) formative assessment includes the kinds of testing instruments used in summative assessment, but, crucially, it includes the kind of contextualised assessment that teachers can provide while the task is being done. Ellis distinguishes between planned and incidental formative assessment. Planned formative assessment requires the use of direct tests of the system-referenced and performance-referenced kinds and must be syllabus-driven. By contrast, incidental formative assessment is "the ad hoc assessment that teachers (and students) carry out as part of the process of performing a task that has been selected for instructional rather than assessment purposes" (Ellis, 2003: p. 314).

Incidental formative assessment is something that results from teacher-learner interaction during or after performing a given FL learning task. During the task the teacher (or peer) can provide online feedback by means of scaffolding strategies (see next section). After the task, the teacher and the learners can reflect on the aspects they noticed.

The kind of formative assessment that seems relevant for an ICALL setting is planned formative assessment. One obvious reason is that CALL is based on computer-learner interaction and is often done in contexts where the teacher does not intervene immediately. A second reason is that programming computers to provide feedback to learners critically requires us to plan what is expected from learners, how it will be assessed, and what aspects should the learners' the attention be drawn to.

4.4 Feedback as a means to help learners

As we noted in Section 4.3, the difficulty of learning a foreign language can be overcome using a concerted strategic approach that includes assessment (Brown, 2007: p. 273–275). Ellis (2003: pp. 180–181) argues that, within task-based instruction, the social dimension of developing a new skill should be related to what socio-constructivists call *scaffolding*. According to Wood et al. (1976), scaffolding includes motivating the learner to perform the task, facilitating the task, maintaining the pursuit of the goal, controlling frustration during the task, marking critical features and discrepancies between learner production and the ideal solution, and demonstrating an idealised version of the act to be performed.

This section focuses on one aspect of scaffolding, namely the marking of critical features and discrepancies between learner production and the ideal solution. Particularly, we describe the types of feedback found in the FLTL literature, and review research on how learners use feedback in CALL contexts.

4.4.1 Types of feedback

Within the rationale of Form-Focused Instruction, a place to implicitly or explicitly draw the learners' attention to language form must be defined. FFI offers a range of possibilities, from explicit metalinguistic explanations to more implicit references to form. Researchers in FLTL and SLA use concepts such as incidental feedback, noticing, or grammar consciousness raising (Brown, 2007: p. 276). Such concepts were developed in studies of teacher-learner interaction in face-to-face instruction, but have a correlate in CALL-based instruction.

Brown (2007: pp. 277–278) presents the different strategies that teachers can rely on to provide learners with feedback as part of form-focused instruction:

1. Recast: It consists in reformulating or expanding an ill-formed or incomplete utterance in an unobtrusive way.

2. Clarification request: It consists in eliciting a reformulation or repetition from a learner, without giving him or her a corrected solution.

3. Metalinguistic feedback: It consists of comments or questions related to the linguistic accuracy of the learner's utterance.

4. Elicitation: It consists in prompting the learner to self-correct without the need of giving him/her the correct version.

5. Explicit correction: It consists in indicating that the learner's utterance is wrong and providing him/her with the correction.

6. Repetition: It consists in repeating the learners utterance with a change in the intonation in the relevant expression or word.

4.4.2 The effectiveness of feedback

The critical aspect of feedback is not only how it can be done, but also what effects it can have. According to Brown (2007: pp. 278–280), research on the effectiveness of form-focused instruction as a means to assist the learner during the learning process "raises more questions than answers". However, he argues, there are two aspects that influence the effectiveness of feedback:

- the ability of the learner to notice a form and its relationship to the feedback, and

- the learner characteristics and style, that is, whether he or she is an analytic or a relational person, field-dependent or field-independent, left or right-brained-oriented, and so on.

Brown (2007: p. 279) argues that it would be useful to know about the optimal time to provide feedback on form, or whether particular linguistic features are more affected by feedback than others, or whether the frequency of input/exposure makes a difference. The research carried out to date does not allow him to draw any firm conclusion. Nonetheless, he suggests that the teacher's task is to "value learners, prize their attempts to communicate, and then provide optimal feedback" for the language system in the learner brain to evolve (Brown, 2007: p. 281).

4.5 Feedback studies in CALL

Within CALL, there are a few specific studies that investigated the effects, use and usefulness of feedback. Two research lines can be identified on the use and usefulness of feedback in CALL (Heift, 2004: p. 417): One investigates the effects of different types of feedback in learning outcome, among which we find Nagata (1993, 1995, and 1997b), and, to a certain extent, Petersen (2010), who compares the learning gain obtained through computer-based instruction with the gain obtained in a face-to-face instruction setting. The second line investigates what learners actually do with feedback: whether they pay attention to it, whether they use it, when they use it, whether they look at sample answers when available, and so on. This second research line is reflected in Pujolà (2001) and Heift (2001a), and to a certain extent in Heift (2004), who studies the correlation between types of feedback and learner uptake.

4.5.1 The effectiveness of feedback in CALL

Nagata's work focuses on the effectiveness of different types of feedback for learning the use of particles by English learners of Japanese in a university language course. We present three of her studies, all of which were performed on English learners of Japanese in the first or second year at university level. Her studies were carried out with student populations ranging from 18 to 34 subjects.

Nagata (1993) compares the effectiveness of traditional feedback with that of metalinguistic feedback provided by an ILTS using NLP. In practice, traditional feedback amounts to reporting to the learner whether his/her response has an unexpected or a missing particle. Intelligent feedback provides the learner with detailed grammatical explanations for the source of errors. Her conclusion is that intelligent feedback is significantly more effective than traditional feedback.

Nagata (1995) presents a new experiment, very similar to Nagata (1993). The main difference is that in this new experiment traditional feedback identifies the position in the sentence where a particle is missing, not simply that there is a particle problem in the sentence. In this study, Nagata concludes that, for learning of grammatical and semantic functions of Japanese particles, intelligent feedback is significantly more effective than enhanced-traditional feedback (Nagata, 1995: pp. 62–64).

Nagata (1997b) investigated whether deductive feedback is more effective than inductive feedback. In her experiment, both types of feedback indicated which particle was wrong and which particle should be used. However, while deductive feedback provided explicit grammatical rules including metalinguistic information at the level of morphsyntax and semantics, inductive feedback provided a set of relevant examples, instead of rules (1997b: p. 524–525). Although she found that deductive feedback is more effective than inductive feedback, the difference was not statistically significant (1997b: p. 530).

4.5.2 Computer-based feedback vs. teacher feedback

Petersen (2010) studied the effectiveness of recast-intensive conversational interaction performed between teachers and learners, and between computers and learners. His student population consisted of 56 subjects. The study focused on the developmental gains in English question formation and morhposyntactic accuracy in young learners of English. One a group received feedback from the teachrs in a face-to-face instruction setting, another received feedback from ICALL system, and a control group received no instruction at all.

The study concluded that recast-intensive conversational interaction facilitated developmental gains in both teacher-learner interaction and in computer-learner interaction (Petersen, 2010: p. 184). Particularly, Petersen concluded that recast-intensive interaction promoted L2 development in ESL question formation and syntactic accuracy in both modalities. Interestingly, only the computer-guided recasts group demonstrated significant gains over the control group in morphological accuracy (2010: p. 184).

4.5.3 The use of CALL feedback by learners

Pujolà (2001) investigated the effects of incremental feedback in reading and listening comprehension activities through multiple-choice or true/false questionnaires. The learner was offered only a quantitative measure in the initial feedback message, and then he or she could decide to ask for more. Pujolà uses the term "immediate feedback" to describe the quantitative measure provided to learners, and he uses the term "delayed feedback" for information provided to learners after they require it per button click. According to him, this presentation strategy adapts to different learner styles: Learners who prefer discovery learning can decide to get more or less information at different points in time, while learners who prefer precise directions can obtain them on demand.

Pujolà (2001) conducted the study with 22 Spanish learners of English who were recorded during the interaction with the materials and who were subsequently interviewed. The study looked at the effects of immediate feedback and the effects of the so-called delayed feedback.

As for immediate feedback, provided during a multiple-choice activity, Pujolà's study found that when learners choose the right responses they do not read the explanation provided for reinforcement purposes (2001: p. 88). By contrast, when learners choose incorrect responses, three different behaviour patterns are observed (2001: p. 87–88):

- Some learners choose another option immediately
- Others re-access the text/audio before choosing another option
- Others think about and select alternative responses

As for the delayed feedback, evaluated using a true-false activity, the author observed that there were mainly two patterns of behaviour, which we interpret as two variations of the same pattern (Pujolà, 2001: p. 88–89). The first of the patterns proceeded as follows: (i) the learner reads the global results stating the number of correct and incorrect answers; (ii) the learner scrolls up to the questions to see which are correct and which incorrect; (iii) the learner requests explanation for a specific question; and (iv) the learner reads the explanation. The second pattern consists of these same four steps, but includes a fifth one. This fifth step involves the learner going back and forth between the explanation and the question to find the source of the error. In other words, the learner repeats steps (iii) and (iv).

4.5.4 The use of ICALL feedback by learners

Heift (2001a) conducted a similar study focusing on the use learners made of error-specific feedback generated by E-Tutor for grammar and vocabulary exercises. The study was conducted on 33 beginner learners of German at university level. The results show that learners reacted in five different ways to system feedback (2001a: p. 103):

- They corrected the error(s) explained by the system.
- They corrected an error in the sentence,but not the one explained by the system.
- They changed a correct structure.
- They resubmitted the same sentence.
- They requested the correct answer(s).

Heift (2001a: p. 107) showed that for the vast majority of sentences students attended to system feedback and corrected only the errors that were highlighted by the system. Heift interprets this as an indicator that students do read the feedback messages. Moreover, Heift (2001a: p. 108) found that students attended to metalinguistic feedback and corrected their output accordingly, even if they had the opportunity to look at the correct answer. These findings indiicate a willingness by learners to be informed about errors.

In a later study, Heift studied the correlation between types of feedback and learner takeup among beginner learners of German at the university level with respect to the following three types of feedback (2004: p. 419):

- Metalinguistic feedback.
- Highlighting[4] and metalinguistic feedback.
- Highlighting and repetition.

[4]Highlighting consists in using a graphical strategy to draw the learner's attention to the location of an error – with a particular font colour or shade.

The study concludes that in those circumstances experiment subjects were more likely to correct their mistakes with the feedback that combined highlighting and metalinguistic deployment, while they were less likely to correct them if the feedback combined highlighting and repetition. Heift found differences with respect to the variables gender and skill level, but these are not statistically significant.

4.6 Chapter summary

In this chapter, we presented fundamental aspects of the research and practice in SLA, FLTL and CALL. In particular, we introduced the concept of Communicative Language Teaching, an approach that emphasises the need of teaching and learning languages as a means for interaction. Within CLT, we focused on Task-Based Language Teaching, a methodology that relies on tasks simulating aspects of real-life communication settings to elicit from learners those linguistic and communicative elements that we expect learners to be competent on.

As we described, the consensus of SLA and FLTL researchers is that a well-justified attention to form, as in FFI, is not only positive but also needed in communicative approaches to language learning – thus, the existence of more language-oriented tasks and more communication-oriented tasks. TBLT theorists and practitioners propose classifying tasks according to their communicative nature. Estaire and Zanón (1994) and Ellis (2003)'s propose distingushing among tasks that purely focus on form and tasks that purely focus on meaning, but Littlewood (2004) proposes a continuum between the two with several stages that allows for a gradual and fine-grained classification. Given that there is no intrinsic goodness or badness in the nature of tasks (Brown, 2007: pp. 18 and 241, Ellis, 2003: Ch. 8), our aim will be to be able to learn the characteristics of the FL learning activities that are suitable for NLP-based assessment, that is, those that are part of the viable processing ground.

We also presented Estaire and Zanón (1994)'s framework for the design of TBLT-driven materials. This framework requires a thorough specification of the pedagogical and linguistic goals of each learning activity foreseen. These specifications provide the information that helps identify relevant linguistic features in the domain of application in terms of NLP.

Another important concept introduced is the assessment of learner production. We presented in detail Bachman and Palmer (1996)'s framework to characterise language tests in terms of target language use setting. Such a framework facilitates a detailed specification of the pedagogical goals and the communicative and linguistic properties of the emulated communicative setting. It includes strategies to describe activity instructions, expected learner responses, and the reciprocal influences that activity instructions have on the expected learner responses, which serve as seeds to not only determine the language to be learnt, but also the language to be processed in an ICALL context. These concepts were suggested as very relevant by Bailey and Meurers (2009) for the characterisation of the viable processing ground and will be further investigated in this thesis.

Finally, we discussed the importance of feedback and feedback types, and a series of studies in CALL and ICALL that investigated the effects of feedback on learning gains, as well as on the use of feedback by learners. The importance of guaranteeing the learner the control over the access to feedback is emphasised by two studies, Pujolà (2001) and Heift (2001b). Moreover, Nagata (1993, 1995), and Heift (2004) found that feedback strategies including metalinguistic feedback increased the performance of learners doing ICALL tasks in form-focused instruction.

Part III

ICALL tasks
Where FLTL meets NLP

[A] well-defined task design with its clear set of relevant language constructions facilitates the restriction to a linguistic domain which is 'manageable' for a system's natural language processing modules.

"Taking Intelligent CALL to the Task"
Task-Based Language Learning and Teaching with Technology
Mathias Schulze (2010: p. 79)

Chapter 5

Methodological considerations

This chapter presents key methodological considerations regarding ICALL instruction settings. We present the different ways in which assessment occurs in ICALL settings as opposed to the way in which it occurs in face-to-face instruction. After that, we explore the relation between activity instructions, learner language, the language processing module and the feedback generation module.

These methodological considerations allow us to focus the object of study of our research. Particularly, we will consider the activity, the learner responses, the language analysis module, and the feedback generation module as four of the elements whose characterisation is critical for the specification of the linguistic properties of the expected learner responses, as well as for their assessment criteria. Such linguistic properties and assessment criteria turn into implementation requirements for the NLP-based feedback generation solutions.

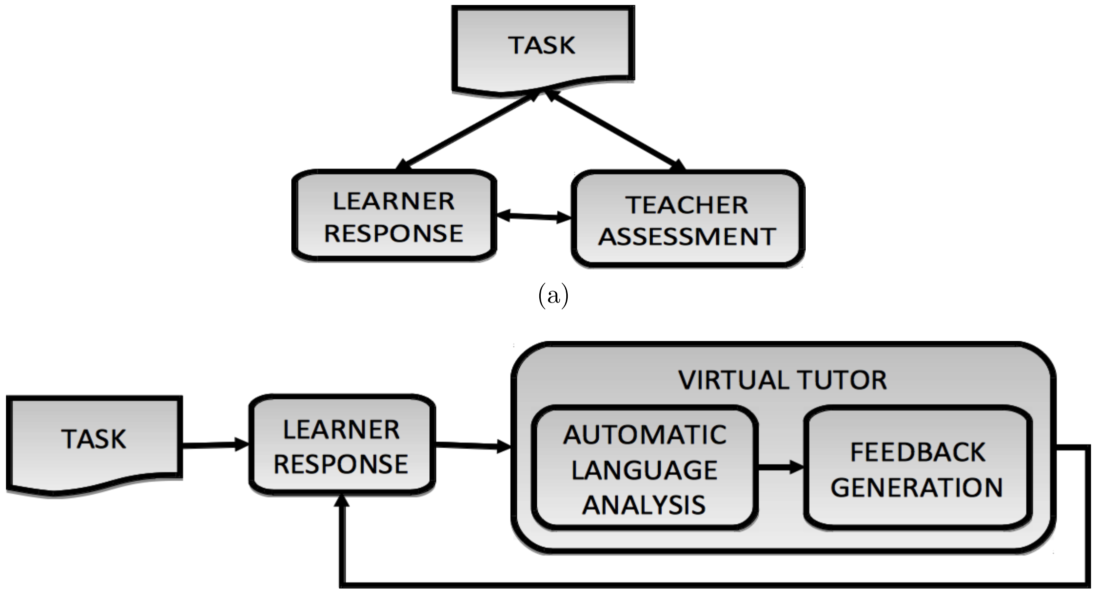
5.1 Teaching and learning in an ICALL setting

As a particular kind of FL learning material, ICALL activities differ crucially from other kinds of learning materials in that, while learners do them, they interact with a virtual tutor enhanced with automatic assessment functionalities.

5.1.1 Interaction flow in an ICALL setting

Figure 5.1 shows two simple graphs reflecting the interaction flow between learner and teacher in face-to-face instruction, Figure 5.1a, and the interaction flow between learner and virtual tutor in ICALL settings, Figure 5.1b.

In both settings the learner is exposed to an activity for which he or she produces a response. As shown in Figure 5.1a, in face-to-face instruction, the teacher provides feedback as a reaction to the learner response. This feedback depends on many of the variables of the learning setting and on the knowledge the teacher has of the learner, all of which is managed by the teacher. The double direction arrows from and to the task pointing to both the teacher and the learner indicate that the task goals or content might be negotiated between teacher and learner.



(b)

Figure 5.1: Differences in teacher/learner and virtual tutor/learner interaction during the learning.

In contrast, as shown in Figure 5.1b, the feedback that can be provided by a virtual tutor depends on the capabilities of the modules that make possible the Automatic Assessment. In our figure, Automatic Assessment includes the NLP-based linguistic analysis module and the feedback generation module, which "interpret" the learner's response, "make assumptions" about its correctness, and "draw" conclusions from as much information as can be modelled. This is a crucial difference that reinforces the importance of anticipation of learner behaviour and learner language as a means to define efficient system reactions.

Using Estaire and Zanón (1994: p. 49)'s phases in the life cycle of a unit of work, we distinguish the design phase, the execution phase, and the evaluation phase. The time in which the learner interacts with the virtual tutor is the execution phase. However, the use of an NLP-enhanced automatic assessment strategy in the execution phase has consequences on the other two phases of the life cycle. We claim that, in order to encompass the pedagogical needs and the capabilities of the NLP tools, a dual perspective, a pedagogical-technological perspective, has to be adopted from the beginning to the end of the cycle.

5.2 The life cycle of ICALL tasks

With this goal to obtain pedagogically meaningful and computationally tractable tasks in mind, we propose a cyclic span of life for ICALL materials to be used in settings combining face-to-face instruction with computer-based instruction. This cycle of life of the ICALL task is presented in Figure 5.2.

Figure 5.2 reflects the processes and interrelationships between Activity, Response, Linguistic Analysis module, and Feedback Generation module during the design and the execution phase. The evaluation phase, at the bottom part, is the last phase of this iterative process: from design to execution, from execution to evaluation, and then back again to design.

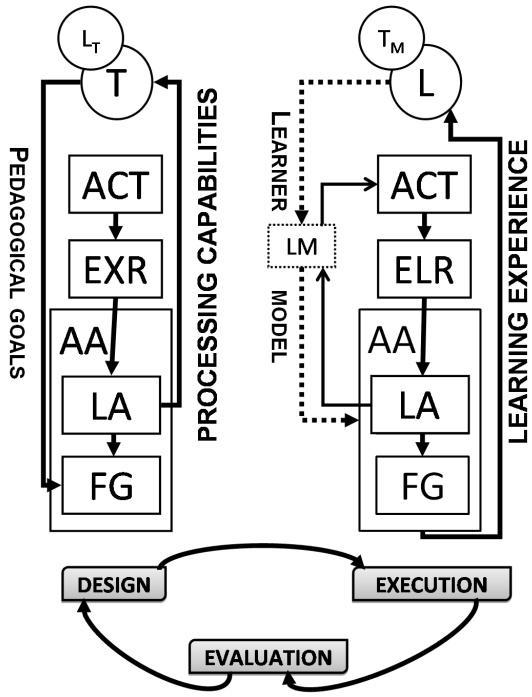


Figure 5.2: The ICALL task life cycle.

In the design phase, we distinguish an activity (ACT), a set of expected responses (EXR), and the automatic assessment (AA) consisting of the linguistic analysis module (LA), and feedback generation (FG). In the execution phase, instead of expected responses we have elicited responses (ELR), that is, actual responses.

Figure 5.2 shows the teacher (T) is on top of the design phase. The teacher is responsible for conceiving and producing the learning materials, and in doing so has the target learner in mind (L_T). In contrast, the learner (L) is on top of the execution phase; the learner performs the activity, and the teacher is in this phase in her/his monitoring role (T_M), which can be performed with the assistance of a virtual tutor, and which is in fact the case in an ICALL setting.

Our approach to encompassing pedagogical needs and NLP limitations is to describe and analyse the different interrelationships that emerge between activity, response, NLP-based analysis and feedback generation. Of course, the interrelationships that we refer to are influenced by the teaching and learning processes, as well as by the participants in the development and execution of the learning activity. However, this latter type of influence falls out of the scope of this thesis, since these are research areas corresponding to more behavioural or cognitive studies.

In the following sections we explore the interaction flow and the relationships between the elements identified in Figure 5.2. We start with the execution phase, because of its central place in the teaching/learning experience. Then we go on with the design and the evaluation phase.

5.2.1 Interaction flow in the execution phase

In the interaction flow during the execution phase several relationships emerge, and they follow a concrete chronological order: there is an activity that is responded to by a learner; the response is analysed with tools for the automatic analysis of language; and, eventually, the automatically analysed learner responses are used by the feedback generation module to provide the learner with assessment. This is a one-way step-by-step process that the learner can start up again after interpreting the system's feedback, as shown in Figure 5.2 by the arrow labelled "learning experience".

Note that the "only" varying element is the learner response. The activity, the linguistic analysis module and the feedback generation strategy do not change over time in the execution phase, in spite of the fact that assessment might be different according to changes in the learner response – unless there is a learner modelling module (see next paragraph). Different learner responses generate different automatic analyses: The feedback generation strategy is dynamic within a range of limited possible forms of behaviour. Automatic assessment is, should be, systematic, that is, feedback messages are repetitive and repeated over a range of learner responses.

The right-hand side representation of the ICALL task's life cycle in Figure 5.2 reflects the existence of a learner modelling (LM) module during the execution phase. Its outgoing and incoming arrows reflect the possibility for the learner model to influence the selection of the FL learning activities chosen for the learner; or the selection of feedback types and strategies to be followed. In this context, the linguistic analysis module provides information to the learner model about linguistic phenomena that can inform about the learner's progress. Nevertheless, learner modelling falls out of the scope of this thesis, indicated by the dotted-lines in its box in Figure 5.2.

5.2.2 Interaction flow in the design phase

As shown on left-hand side of Figure 5.2, the starting element in the design phase is the activity, produced by the teacher – or a content creator. During this phase, the four elements of the ICALL activity maintain their one-way relationships described in the execution phase, but other non-linear relationships emerge.

The first relation in the "straightforward" information flow emerges between the activity conceived and a range of expected responses. The teacher, with a target learner in mind, identifies a set of pedagogical needs related to specific communicative and linguistic skills. The activity results in a focused task (Ellis, 2003: p. 16–17), in which the wording of the instructions and the means given or pointed to to learners are oriented to the practising of the targeted skills. These skills are related to particular linguistic structures to be elicited on the learner side, which might allow for the specification of a range of expected responses – in other words an NLP domain.

We identify a second relationship between the expected responses and the linguistic analysis module, one that connects the contents and the language of the expected responses with the linguistic analysis module. The language analysis module requires a fine-grained specification of the lexical elements that need to be in the response, as well as the corresponding linguistic relations: fonetic, morphosyntactic, semantic or pragmatic.

This relationship between the expected responses and the linguistic analysis module is influenced by pedagogical considerations such as the pedagogical goals, the instructions, or the input data. Thus, if an activity targets at training learners on the use of a particular syntactic structure, an automatic analysis module that provides the appropriate syntactic analysis is required. If the activity targets at practising the use of writing abilities such as expressing interest for a job, then pragmatics, the functional contents in pedagogical terms, have to be correspondingly modelled.

The third straightforward relationship is the one between the linguistic analysis module and the feedback generation module. Whatever it has to be *said* by the virtual tutor it has to be based on evidence found in the linguistic material that is present or absent in the learner's response. Thus, the capabilities of the software for the automatic analysis of language play a key role in the appropriate detection of the expected language and contents in the elicited learner responses.

As for the non-straightforward interrelationships, we identify two of them. On the one side, the pedagogical goals of the activity might affect the feedback generation strategy; on the other side, the capabilities of the language processing tools, known by the NLP developer, might affect the pedagogical design.

The limitations of the NLP tools determine the kinds of ICALL activities that can be successfully implemented. To put a simple example, a semantic analysis module that works only at the sentence level will not be enough to assess texts containing more than one sentence. In such a case, the NLP tools might be enhanced to work beyond the sentence level, or one might decide to rethink the pedagogical concept.

The second non-straightforward relationship emerges between the pedagogical goals and the feedback generation strategy. If the activity focuses on form or on meaning, or on specific aspects of form, different kinds of linguistic or communicative issues will be prioritised as part of the feedback. Similarly, depending on the purpose of the assessment (low, medium or high-stakes), the feedback generation process will require different levels of linguistic information and different types of post-processing of the linguistic information. Last but not least, the desired nature of feedback has an impact too. To provide both positive and negative feedback, a module for the automatic analysis of learner responses requires analysing both correct and incorrect elements in the response.

5.2.3 Interrelationships in the evaluation phase

In the evaluation phase all the interrelationships considered in the two previous phases have to be re-visited. At this stage, the activity has been performed by learners, assessed by the virtual tutor, monitored by the teacher, and can be evaluated in terms of success. The goal of the evaluation is to validate the activity as one that helps teachers and learners accomplish their respective goals. By comparing the results obtained to the objectives initially defined, a set of recommendations regarding changes and improvements can be made.

From our perspective, there are mainly three aspects that need to be looked at in the evaluation of an ICALL activity: (i) whether the activity is pushing the learner to practice the targeted communicative and linguistic skills; (ii) whether the learner is capable of improving its outcome with the help of the automatically generated feedback; and (iii) whether the performance of the feedback generation module is undermined by the performance of the NLP tools. The first one is not specific of an ICALL setting, but the second and the third are.

If learner outcomes correspond with those that favour the acquisition of the targeted skills, the activity is accomplishing its goal. If not, there are at least two important questions to be considered. First, whether there is a flaw in the design or in the execution of the activity that prevents the learner from achieving the expected pedagogical goals. Second, whether unexpected or incoherent behaviours in the NLP-based correction functionalities can be attributed to the incapability of the NLP system to adjust to linguistic structures different from the expected.

If the learner is incapable of improving her/his learning outcomes, then the feedback is not achieving its goals.[1] The causes might be on the learner side, because s/he might not be paying attention to it, or not noticing it. Or on the virtual tutor side (the ICALL design team), because the feedback strategy chosen does not correspond to or is not compatible with the learner's style, background or level.

Finally, if the performance of the feedback generation strategy is flawed by the performance of the NLP tools, then critical inconsistencies or misleading feedback messages might show up. For instance, the feedback generation strategy will not be reliable in the identification of the use of the definite/indefinite determiner if either the analysis or error detection modules for that linguistic phenomenon do not perform with the required precision and recall.

The comparison between expected performance and actual performance will inform of the changes and improvements to be made in the ICALL materials from its pedagogical conception to its use "in class", through the computational implementation of the NLP resources, its graphical presentation, and so on.

5.3 Connecting FLTL and NLP in the lifecycle of ICALL tasks

Figure 5.3 presents in different coloured boxes the elements of the life cycle of an ICALL task that we focus on in this part of the thesis. The Activity and the Expected Response are highlighted in blue in the figure, and these two entities of the ICALL task correspond with two methodological instruments we present. To determine the needs and characterise the Activity we propose the Task Analysis Framework (TAF), inspired by the task characterisation and classification criteria reviewed in Chapter 4. To produce instruments to specify Expected Responses, and the assessment criteria, we propose the Response Interpretation Framework (RIF), inspired by the test task characterisation framework presented in Chapter 4. Both are introduced in and exemplified in Chapter 7.

The Linguistic Analysis module and the Feedback Generation module are highlighted in green in Figure 5.3. These two elements of the ICALL task are the target of Automatic Assessment Specification Framework (AASF), a framework to work

[1]Assuming learners with normal cognitive skills and an activity that is appropriate for his/her profile.

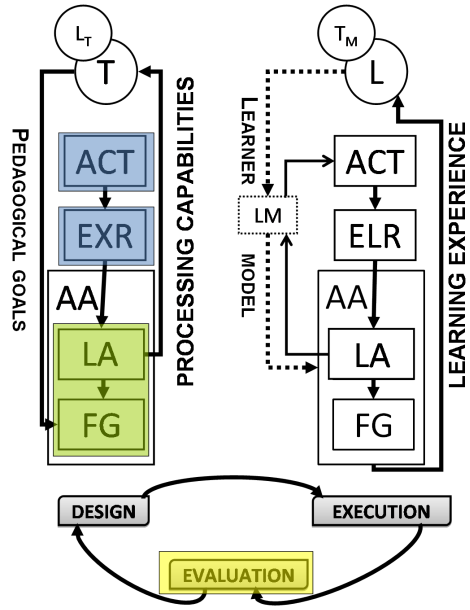


Figure 5.3: Processes and interrelationships within the ICALL activity focused on for the definition and exemplification of the methodology proposed.

out the design and implementation of the Linguistic Analysis module and the Feedback Generation module. The AASF is a means to implement NLP-based automatic assessment modules taking as input the pedagogical and linguistic specifications generated on the basis of the TAF and the RIF. Moreover, it determines the implementation of assessment strategies to provide learners with formative and/or summative assessment. The AASF is presented and exemplified in Chapter 8.

Finally, Figure 5.3 highlights the evaluation phase in yellow. In this phase, the differences between the expected and the elicited responses can be examined, as well as the consequences these differences have on the pedagogical design and system behaviour. In Chapter 9 we present an empiricial analysis of some of the learner responses collected for the ICALL tasks exemplified in Chapters 7 and8.

5.4 Chapter summary

In this chapter we introduced key methodological considerations regarding the nature of ICALL instruction for the purpose of this thesis. Typically, while learners use ICALL materials, teachers do not provide immediate assistance. This requires the anticipation of learner behaviour and of system use. We focused our research on the analysis and characterisation of the relationships existing between activities, the language they are expected to elicit from learners, the module for the automatic analysis of learner language, and the module for the generation of feedback messages.

This focus corresponds to the contents of Chapters 7 and 8, where we present methodological instruments to facilitate the pedagogically-driven characterisation of tasks and learner responses as an input to the specification of computational needs.

Chapter 6

A research setting to develop ICALL materials

This chapter presents the pedagogical and technical setting that we will use to exemplify the different methodological instruments that we propose for the development of ICALL materials. This setting is derived from ALLES[1], an EU-funded project, whose goal was the design, implementation and evaluation of distance language learning materials including NLP-based correction functionalities for four different languages.

In pedagogical terms, our instruction setting followed a task-based approach and, in particular, we followed Estaire and Zanón (1994)'s framework for the development of task-based syllabi. In computational terms, our development setting made use of pre-existing shallow-processing robust NLP tools.

6.1 Overall perspective

ALLES, the name of the project that we use as a research setting, is an acronym that stands for Advanced Long-distance Language Education System.

The project's main goals were to prove the concept that:

1. Task-Based Language Instruction could be implemented and used for computer-assisted self-learning; and that

2. Natural Language Processing techniques could help develop CALL materials with intelligent tutoring capabilities in line with communicative approaches to language teaching.

As reflected in Table 6.1, the ALLES consortium consisted of five partners: Atos Origin Spain, Fundació Barcelona Media Universitat Pompeu Fabra, Heriot-Watt University, the Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung e.V. an der Universität des Saarlandes, and Universidad Europea de

[1]This chapter is based on project work carried out by a group of researchers and technicians in the ALLES project, with whom the I was fortunate to collaborate. ALLES was a three-year project funded by the European Commission under the 5th Framework Programme (contract number IST-2001-34246).

Institution name	Location	Role in the project
Atos Origin	Madrid	Coordination Interface design Lexical resource development
Fundació Barcelona Media Universitat Pompeu Fabra	Barcelona	FLTL expertise NLP expertise Architecture design and implementation FLTL and SLA expertise FL teaching in Catalan, English and Spanish
Heriot-Watt University	Edinburgh	FLTL and SLA expertise FL teaching in English and German
Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung e.V. an der Universität des Saarlandes	Saarbrücken	NLP expertise FL teaching in English in Universität des Saarlandes
Universidad Europea de Madrid	Madrid	FLTL expertise FL teaching in English

Table 6.1: Partners of the ALLES consortium and their respective expertise.

Madrid. As the table shows, the project involved the collaboration of experts in several fields, mainly in Second Language Acquisition, Foreign Language Teaching and Training, Computer Science (Software Engineers and Graphical Designers), Linguists and Computational Linguists. There was a total of 26 people involved in the project, with a permanent staff of six to eight people. These figures show the complexity and variety required in teams aiming at the development of ICALL systems.

As a Computational Linguist, my role in the project was to design and develop pedagogically informed NLP tools for the generation of automatic feedback. Particularly I worked in the design and implementation of surface shallow semantic processing techniques for the automatic evaluation of learner responses, in the design and implementation of summative assessment strategies based on automatically annotated text, and in the specification of expected responses with the collaboration of content designers. Though I specifically worked on the development of NLP resources for Catalan and Spanish, the design work was language independent and affected the four languages of the project.

The ALLES consortium eventually designed and developed web-based materials for Catalan, English, German and Spanish including NLP-based automatic correction facilities. ALLES materials were trialled between June 2002 and May 2005 by second language teachers in four different universities with different degrees of implication and success. The materials are available under http://www.iai-sb.de/alles. ALLES is not being used in a continued manner in real-wolrd instruction settings, but a subset of the materials was used in a project called AutoLearn (Estrada et al., 2009).

As reflected in the project's final report, within ALLES we managed to (Martín et al., 2005: pp. 11–23):

- Develop a set of CALL materials that are in line with the principles of Communicative Language Teaching following the task-based instruction approach.

- Integrate NLP-based immediate individualised feedback in an e-learning platform to be used for self-learning.

- Prove the concept that formative and summative assessment generation strategies in line with communicative approaches to language instruction can be implemented on the basis of NLP techniques.

The chapters to come in this part of the thesis describe how formative and summative assessment strategies could be developed in that setting by integrating FLTL and NLP insights.

6.2 TBLT-driven design of materials

ALLES materials were designed for the instruction of Language for Specific Purposes – targeting learners of Catalan, English, German and Spanish as a foreign language in the business and finance. As for the level of proficiency of target learners, ALLES targeted at B2 level or the C1 level learners as defined by the Common European Framework (CEF, of Europe, 2001).

In this section we describe how the first steps in the planing phase of Estaire and Zanón's framework (see Section 4.2.2) were used in ALLES to obtain concrete pedagogical specifications for the development of materials. These specifications were a first step into the characterisation of the communicative and linguistic properties of the learning activities.

6.2.1 Determining an interest area

The first step in Estaire and Zanón's framework is to determine a list of interest areas. ALLES content designers relied on their previous experience and on a review of available materials for the learning of the relevant languages in the business domain (Díaz, Ruggia, and Quixal, 2003a: p. 4). Figure 6.1 shows the resulting list of interest areas per level independently of the language.

As shown in Figure 6.1, each interest area has a different topic for each of the proficiency levels for which contents were developed. The columns "B2 Level" and "C1 Level" show the topic names for the corresponding CEF levels. At this level, content design is language independent.

In the coming sections, we detail the application of the next five steps in the planning phase of Estaire and Zanón (1994)'s framework. We exemplify it in one of the ALLES learning units, namely for *Education and Training*, the B2 level topic within the interest area *Career Management and Human Resources*. See the complete description of the two learning units in the interest area *Career Management and Human Resources* in Appendix A.

Interest Areas	**B2 Level**	**C1 Level**
Career Management and Human Resources	Education and Training	Job Interview (describing jobs responsibilities)
Business Communication	Customer Service, International Communication	Business Reports
Management and Organizational Behaviour	Company Organization	Planning and Control
Financial Aspects	Personal Finance (facts, and figures, loans, credits, cards, banking)	Corporate Finance
Marketing	Product Benchmarking	Advertising
Stock Exchange	Customer Portfolio (company report)	National and International Markets

Figure 6.1: ALLES topics according to interest area and learner CEF level taken from Díaz et al. (2003a: p. 4).

6.2.2 Planning a final task

The second step of Estaire and Zanón's work plan is to devise a final task. For the learning unit *Education and Training*, the specifications of the final task read (Díaz, Ruggia, Quixal, Torrejón, Jiménez, Rico, Garnier, and Schmidt, 2003b: p. 6):

> At the end of the unit, the student will write an email where he will register for a training course offered at his company. In this email the student will specify reasons why he is interested in taking this course and the timetable. To complete this task, the student will use:
>
> 1. The course listing attached by Human Resources to the email describing the availability of training courses
> 2. His schedule for the current month
> 3. Voice mail from his boss recommending a particular course

In this task, content designers planned a role play activity in which an employee is expected to send an email to register for a course offered by the human resources department in a fictive company. This description indicates the communicative skills and linguistic products implied in the setting: writing emails, understanding oral recorded messages, understanding course descriptions, etc. Moreover, some input data is provided to the learner: a month schedule, a message from the manager and a list of the available courses. This input data has to be processed by the learner to complete the task.

6.2.3 Determine the unit objectives

With the **final task** in mind, ALLES content designers define the unit objectives, which are related to rather general communicative and linguistic skills. What makes them specific is the fact that they are contained in a specific topic-determined pedagogical setting. For the unit *Education and Training*, the unit objectives are (Díaz et al., 2003b: p. 6):

> During the unit the students will develop, with a degree of communicative competence in accordance with their level, the ability and knowledge necessary to:
>
> - Understand requirements to register for courses.
> - Write emails in order to complete a registration.
> - Speak about her or his interests.
> - Know how to write professional emails (structure, expressions, tone, etc.).

These are the communicative skills in which learners are expected to gain competence by going through the learning unit. Some of them are specifically addressed in the final task, for instance, the first two in the list, but others might be practised in previous preparatory tasks.

6.2.4 Content specification of the unit of work

As for the three main types of contents foreseen by Estaire and Zanón, the ones specified by ALLES content designers in the unit *Education and Training* are (Díaz et al., 2003b: p. 6):

- Thematic content
  - Registration process for in-house training courses
  - Professional emails
  - Motivation of workforce[2]
  - Corporate training courses[2]
- Linguistic content
  - Lexical: words, expressions and gambits used for registration, courses and schedules.
  - Functional content: expressing likes and dislikes, making suggestions, writing an email (techniques, structure, control, ...), recommending and asking for advice, describing (courses).

[2]For the sake of completeness, we add these two thematic objectives, derived from the texts included in the final version of the English unit, though not included in the original version of the unit design.

  - Grammar content: structures used for making suggestions, recommendations, asking for advice, describing things.
  - Textual types: registration forms and e-mails

- Socio-cultural content
  - It will fit the material collected for this unit

The **thematic content** of the unit informs of the broad topics that are expected to be part of the unit. The **linguistic content** informs of the linguistic structures and pieces that are expected to be understood or produced by learners, and it is in turn divided into four further subtypes: **lexical content**, "words, expressions and gambits used for registration, courses and schedules"; **functional content**, "expressing likes and dislikes, making suggestions, writing an email (techniques, structure, control...), recommending and asking for advice, describing (courses)"; **grammar content** "grammar structures used for making suggestions, recommendations, asking for advice, describing things"; and **textual types** "registration forms and e-mails".

Note that three of the subtypes of linguistic content are strictly related with the formal aspects of language: lexical content to vocabulary, grammar content to morphosyntax and sentence structure, and textual content to pragmatics (or text linguistics). The fourth one, functional content, is related to a functional description of language very common in the FLTL field, which is related to the kinds of communicative functions that can be performed with specific linguistic structures. Functional contents are often related to formulae or exponents of function (Estaire and Zanón, 1994: pp. 30 and 58).

Finally, **socio-cultural content** is defined as those social and cultural aspects emerging from the texts and the settings that learners are put in. Socio-cultural content is usually made explicit by requiring certain conventions in the linguistic products expected in each activity.

6.2.5 Process plan

The fifth step in Estaire and Zanón's framework consists in preparing a sequence of preparatory tasks that equip the learner with the knowledge and the competence to succeed in the final task. The *Education and Training* learning unit consists of three tasks (*subtasks* in ALLES terminology) in addition to the final task. Each of the subtasks is expected to help learners develop part of the targeted linguistic and communicative skills.

The corresponding section in the specifications of the learning unit reads (Díaz et al., 2003b: p. 7):

- Subtask 1 (main skill: reading): The student will read a business article regarding the importance of having a properly trained workforce and value of human capital in the companies. Next, the student will read various work schedules from different employees in a company, their job profiles and a list of specialised courses offered by the Human Resources department. They have to match the employees' schedules and profiles with the courses they could take for further advancement in their careers and explaining why these matches are appropriate.

- Subtask 2 (main skill: writing; other skills: listening, reading): The student will listen to a recording of an informal talk between two employees exchanging views on different training courses offered at their company and discussing pros and cons. Next, the student will read some short articles on the use of emails in business settings and how to write formal and informal emails. Finally, the student will write a short informal email to a friend. The email topic will be a description of courses listed on a leaflet and questions about what courses to take.
- Subtask 3 (main skill: speaking): The student will do a role-play activity in which they will call the human resources department asking for seat availability for a particular course, use of laptop during the course, material required, and whether there will be a diploma issued at the end.

At this stage of design, it might still be undecided which of the activities that will be given to learners are *communicative* or *enabling* tasks. This can be decided later on during the actual development of the activities.

6.2.5.1 Learning sequences in ALLES

One additional aspect to be taken into account is that ALLES materials were organised according to three main concepts: learning unit, subtask, and activity (Díaz et al., 2003a, and Díaz, Ruggia, Quixal, Torrejón, Jiménez, Rico, Garnier, and Schmidt, 2004). These three structural concepts determine the way learning materials are presented to learners. The corresponding definitions are (Díaz et al., 2004: pp. 6–8):

Learning unit A learning unit is a structured piece of work consisting of a series of problem-solving subtasks around a topic. The language learning objective is to develop the learners ability and knowledge to do something in the foreign language.

Subtask A subtask is a problem-solving learning work aiming to improve language use or communicative competences. Subtasks are divided into two classes:

1. Communicative subtasks, which might involve any of the four skills and are mainly focused on meaning (rather than form). Final tasks are also communicative tasks in which a variety of competences and skills are required for the learner to fulfil the task.
2. Enabling tasks: tasks where learners practice language possibly with some attention to meaning, but not requiring to communicate new messages.

Activity Activities are the smallest units of work. Activities (one or more of them) correspond to a FL activity in the sense of Ellis (2003: p. 15).

This hierarchical structure of ALLES materials responds to the fact that they were conceived as CALL web-based materials with structured activity sequences. Subtasks are organised in series of Activities, which are the learning object in which NLP-based automatic assessment might be integrated.

6.2.6 Instruments and procedure for evaluation

The sixth step in Estaire and Zanón's content development framework is the one they describe as "plan instruments and procedure for evaluation of process and product". Evaluation in ALLES is related to the way activity sequencing is conceived, namely as a series of preparatory tasks leading to a final task.

For preparatory tasks, content designers required formative assessment, since their purpose is to help the learner gain competence to succeed in a communicative setting mirrored in the final task. Formative assessment is part of the so-called scaffolding in communicative language teaching, and is part of the evaluation of the process. As for final tasks, learners are evaluated by requiring them to produce a communicative outcome, which is evaluated following a summative assessment strategy that takes into account both quantitative and qualitative aspects of the product. The definition of formative and summative assessment in ALLES is a product resulting from the integration of FLTL and NLP insights.

6.2.6.1 Formative assessment in ALLES

In ALLES, formative assessment is conceived "as part of the monitoring process" (Badia, Díaz, Garnier, Lucha, Martinez, Quixal, Ruggia, and Schmidt, 2005: p. 6): different types of feedback inform the learner on how well s/he performed on a given task on the basis of the assessment criteria specified.

When a learner provides a response, the different types of feedback foreseen are:

1. Inform the learner whether the response is correct or not.
2. If the response is correct the system provides:
   - Information on (persistent) topical knowledge errors, if any;
   - Information on (persistent) linguistic knowledge errors, if any; and
   - A warning against unnecessary or unexpected information, if any.
3. If the response is incorrect, for each detected error the system provides:
   - The location of the error, unless it is a "global" error or not related to a particular location in the learner response;
   - The explanation of the error; and
   - Possible ways of repairing the error.

This characterisation of formative feedback informs the design of the feedback generation functionalities of the ICALL system. The design is based on pedagogical needs and requirements, which will be realised by exploiting the information generated by the automatic linguistic analysis modules via insights from studies on the nature and effects of feedback.

6.2.6.2 Summative assessment in ALLES

Summative assessment in ALLES is conceived to provide learners with an idea of how effective a product generated by them would be in a communicative setting as the one emulated in the final task. This is always in form of a grade that takes into account four different parameters. The proposal is based on a series of linguistically and pedagogically motivated scoring measures that take into account the activity goals in terms of the information to be communicated – communicative contents – and measures of complexity, accuracy and fluency.

The qualitative/quantitative criteria required by FLTL experts for the assessment of final tasks are:

- Communicative contents: number of functions related to informative contents – listed in the response criteria for correctness – and number of functions related to language knowledge at the level of pragmatics or the level of the text genre.
- Lexical contents: total number of words and word-sentence ratio, and total number of domain-specific words (specific vocabulary).
- Sentence structure and accuracy: simple and complex sentence ratio, number of discourse markers, and number of grammar or word usage errors.
- Overall text layout: number of paragraphs and number of spelling errors.

These measures are based on research carried out by Díaz and Ruggia (2004), which is inspired by (Wolf-Quintero et al., 1998). The challenge for NLP developers is to provide a battery of numeric cues obtained from the automatically analysed version of the learner's response that relate to the above criteria – see Chapter 8.

6.2.7 From the design to the actual materials

The level of detail of the above specifications is still far away from the actual learning materials. This was a conscious decision during the ALLES project, one that was oriented to facilitate two apparently contradictory goals: (i) the creation of a language-independent topic-based syllabus, and (ii) the creation of language-specific and culture-specific materials. By keeping the initial syllabus design at this level, ALLES material developers did not compromise language-dependent aspects until the actual activity was worked out.

The work that follows these specifications down to the actual creation of the learning materials including its automatic assessment functionalities is precisely the focus of our research. The interaction and collaboration between FLTL and NLP experts is what facilitates the development of pedagogically sound materials that can be assessed with NLP-based automatic assessment tools. This interaction benefits from being coupled with a top-down bottom-up material development process: On the one hand, Estaire and Zanón's design framework provided this initial top-down vision of the contents. On the other hand, as the actual materials are developed and found (some times adapted and some times created) unit objectives are further specified or simply re-defined.

In ALLES, the top-down direction of the content development process made it possible to have similar materials for the four different languages. At the same time, the initial design already restricted the type of communicative settings and the topic domain of the materials. The bottom-up direction made it possible to introduce language-specific (or topic or culture specific) features to each learning unit as materials were actually developed.

During the ALLES project the negotiations between FLTL and NLP experts was neither explicitly formalised nor documented in detail. Most of the interesting "work" took place in multi-party meetings, in scattered casual conversations, or in long intense group work sessions in front of the computer. The methodological instruments that we present in the following chapter systematise these negotiations a posteriori.

6.3 A general architecture for the analysis of learner language

This section describes the technical characteristics of the NLP tools that underlie the research presented in Chapters 7 and 8. We present a modular architecture implementing a rule-based approach to linguistic processing with a set of modules that are domain-independent and a set of modules that are domain-dependent. The domain-dependent modules provide the ability to analyse language taking into account some syntactic, semantic and pragmatic properties that are relevant for the activity being assessed – where each activity is taken as a domain of its own as defined in Section 3.1.2.

Figure 6.2 shows the modules of an NLP architecture that provides with spell and grammar checking functionalities and with so-called information extraction functionalities. The figure reflects the difference between domain-independent NLP modules and resources, whose borders present plain lines, and domain-adapted NLP modules and resources, whose borders present dashed lines. In line with Basili and Zanzotto (2002: p. 97–99), our approach relies on modularisation and the adaptability of domain components as the key to robustness.

The three initial modules are the Tokeniser, the Morphological Analyser, and the Morphosyntactic Disambiguator, and these are common to both spell and grammar checking and information extraction. After these, the learner response can be sent to the Non-Word Spell Checker and the Context-Sensitive Spell Checker, or, for the analysis of activity-specific contents to the Information Extraction module.[3]

[3]Chapter 8 describes the actual feedback generation as a two-step process: The first correction step focuses on the correction of formal errors, and the second one on the assessment of the response in global terms.

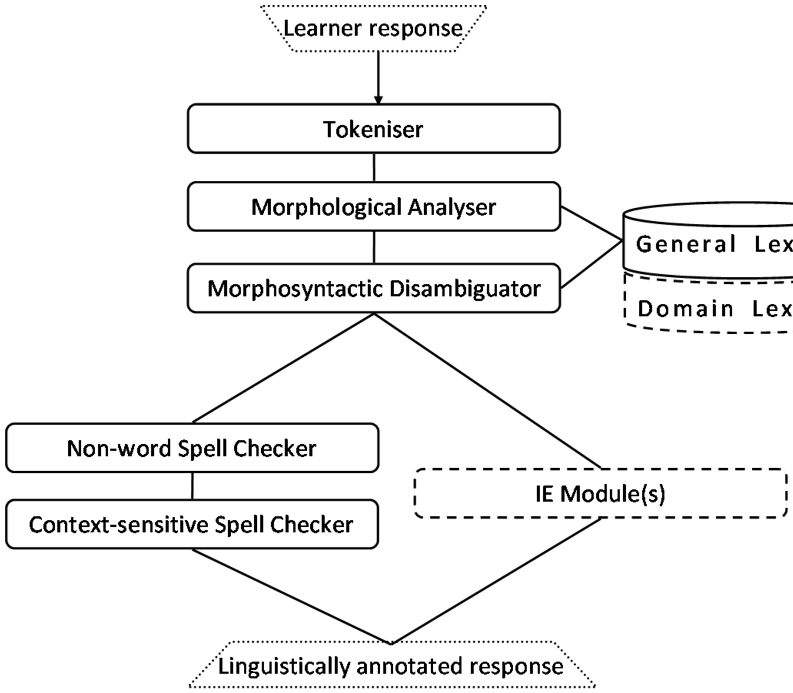


Figure 6.2: A modular and domain-adaptive NLP architecture with for the processing of learner responses.

We describe in further detail the modules of the architecture:

- TOKENISER: It segments text into tokens, mainly words, but it handles textual objects such as numbers, punctuation signs, and so on, as well as sentences and paragraphs. It also identifies other characteristics such as word-case, number of tokens in a sentence, and so on.

- MORPHOLOGICAL ANALYSER: It assigns the corresponding reading(s) to each of the tokens identified. The process might include both dictionary look-up and on-line morphological analysis, depending on the language, or on whether the word under analysis is found in the dictionary. The dictionary look-up might take into account general and domain-specific lexica. The module assigns default grammatical categories and features to those unknown elements on the basis of heuristics. For instance, a word ending in *-tion* or *-ness* is very probably a singular noun, while a word ending in *-ed* is more probably the past participle form of verb.

- MORPHOSYNTACTIC DISAMBIGUATOR: For each word with more than a reading at a morphosyntactic level the most plausible reading is chosen. The decision is taken on the basis of the local context, that is, taking into account the grammatical features and the distribution of a set of words close to each other.

- NON-WORD SPELL CHECKER: It generates a list of correction proposals for each of the tokens not found in the dictionary. To filter the list of generated alternatives the general and the domain-specific lexicon are taken into account.

- **Context-sensitive Spell Checker**: It detects errors resulting into words, that is, errors that cannot be detected by the Spell Checker because the words that constitute them are in the dictionary. In order to detect them the surrounding context has to be taken into account. If it is able to, it generates correction proposals.

- **Information Extraction Module**: This module consists of several submodules. Each of them identifies sequences of linguistic elements that correspond to information chunks required in a particular response. Information Extraction amounts to parcelling meaning or complex linguistic structures into units that can be then used to check for the correctness of the learner's response.

- **General and domain lexicon**: This is a resource including a dictionary containing linguistic information associated with each entry. We assume a dictionary containing for each word all its possible readings, the associated lemmata, and the associated morphosyntactic informations such as number and gender for determiners, adjectives, and nouns, or mode, tense, person, number for verbs, and so on. The domain lexicon will be enhanced with words that are relevant for a particular activity containing the same kind of information.

Chapter 8 describes how the feasibility of an ICALL activity depends largely on the capability of the Information Extraction modules to annotate learner responses with the linguistic information relevant for their assessment in pedagogical terms.

6.3.1 The linguistic analysis underlying domain-specific assessment

The results of the analysis of the first five modules are described in the following paragraphs. The type of linguistic information that these modules provide is the basis for the activity-specific analysis modules, those that provide the adaptivity of the system.

Table 6.2 presents the analysis resulting from the first five modules for the sentence **how satisfied is you witth Stanley Broadband?*. The first column in Table 6.2 is a token identification number. The second column is the result of segmenting the sentence into tokens and identifying sentence boundaries by applying the Tokeniser.[4]

The third column shows the readings assigned to each word by the Morphological Analyser. If a word cannot be processed by any of those two strategies then it is assigned any of three possible readings – noun, adjective or verb, or all of them if no better decision can be made. Guessed readings are correspondingly marked – see *witth*, token no. 5, in the first column, whose readings end with a question mark.

The Morphological Analyser identifies also particular kinds of entities that might be relevant for later states in the processing, if the corresponding information is

[4]The way we represent the information does not necessarily reflect the way the data are internally structured.

included in the dictionaries. This is the case of *Stanley Broadband*, token no. 6 in the third column, identified as an entity of the type product.[5]

The fourth column contains the result of the disambiguation process for those words with multiple readings after the morphological analysis. For instance, in token no. 1 the arrow (⇐) points to the pronoun reading, discarding the adverb reading. In token no. 4 the arrow points second-person singular pronoun reading, discarding the second-person plural reading.

ID	Token	Reading(s)	POS	Error correction
	<s>			
1	*how*	how: pron how: adv	⇐	SentStartCap: How
2	*satisfied*	satisfy: verb part past	⇐	–
3	*is*	be: verb pres ind 3rd pers	⇐	AgrErrorStart: *are*
4	*you*	you: pron 2pers sg you: pron 2pers pl	⇐	AgrErrorEnd
5	*witth*	witth: noun? witth: adj? witth: verb?	⇐	SpellErr: *with, width,* ...
	<entity type=product>			
6	*Stanley Broadband*	Stanley Broadband: noun sg	⇐	–
	< /entity>			
7	*?*	?: punct	⇐	
	< /s>			

Table 6.2: Sentence annotated with the initial levels of analysis of the proposed architecture.

The fifth column in Table 6.2 presents the results of the spell checking and the context-sensitive spell checking modules. The Spell Checker generates a list of correction proposals for unknown words, as is the case of token no. 5 (*witth*). In this same column there are also two examples of context-sensitive spell checking: Token no. 1 is marked as containing a non-capitalisation error (a word starting a sentence is usually capitalised: SentStartCap) and the capitalised word as a correction proposal. Tokens no. 3 and 4 identify a subject-predicate agreement error (codes *AgrErrStart* and *AgrErrorEnd*) and a correction proposal for token no. 3.

6.3.2 Two concrete implementations

During the ALLES project, two different implementations of this architecture were implemented. We used a solution based on two formalisms for the processing of English, German and Spanish text; the formalisms are MPRO (Maas, 1996) and KURD (Carl and Schmidt-Wigger, 1998). A different solution based on the Con-

[5]Entity recognition is often performed in a separate module, particularly if it includes identification procedures that go beyond checking a list of lexical entries, but this is not relevant here.

straint Grammar (known as CG, Karlsson et al., 1995) was used for the processing of Catalan text (Badia et al., 2001).

Both implementations were based on previously existing NLP resources. The MPRO-KURD solution is an application that has evolved for more than 25 years now, and that it is being used, among others, for the analysis of text in domain-independent and domain-dependent tasks such spell and grammar checking of unrestricted text (Verlag, 2010[6]), style checkers in linguistic documentation (Haller, 1996, 2001), machine-translation systems (Streiter and Schmidt-Wigger, 1995), and systems combining machine translation and translation memories (Carl et al., 2002).

The CG-based solution was initially developed as a morphosyntactic (Badia et al., 2001) tagger and was at that time starting to be developed as a spell and grammar checker for native-speakers (Badia et al., 2004; Aguilar et al., 2004) for Catalan. It was used also as a basis for sentence compression tasks (Bouayad-Agha et al., 2006), and for quantitative linguistics studies (Mayol et al., 2005; Boleda, 2007).

Both the MPRO-KURD solution and the CG-based solution have similar approaches to language processing. They are both surface-based shallow analysis tools, they are highly-dependent on hand-crafted rule-based grammars and their underlying computational implementation is finite-state techniques (see Appendix B).

6.3.2.1 The MPRO-KURD solution

The MPRO-KURD NLP processing software consists of several modules based on MPRO and KURD. MPRO (Morphological PROcessing) is a formalism specifically designed for the tasks of tokenisation, morphological analysis and to some extent for the disambiguation of morphosyntactic ambiguities (Maas, 1996; Garnier et al., 2003a). MPRO can be considered a tool able to process unrestricted text. According to Garnier et al. (2003a: p. 8) the German version of MPRO contains 90,000 morphemes, very much corresponding to the entries included in a general dictionary. The Tokeniser, the Morphological Analyser and the Spell Checker are implemented in MPRO, which includes an algorithm based on minimum edit distance measures for the generation of correction proposals – see Jurafsky and Martin (2009: Ch. 3) for a description of minimum edit distance algorithms. KURD-based grammars are used in the Morphological Disambiguator, the Context-Sensitive Spell Checker, and the Information Extraction module.

MPRO modules receive as input a string of text and give as output a linguistic analysis that consists of sentences and tokens. A sentence consists of several tokens, most of them words, and each word is associated with (at least) one feature bundle. A feature bundle consists of attribute-value pairs that provide linguistic information related to the word and to the position and graphical representation of the word in the original text. This process is realised by two modules called LESEN and PARSER, which are respectively responsible for tokenisation and morphological analysis.

Figure 6.3 reflects (part of) the analysis that MPRO yields for the German sentence in (1), which is later on used as input for KURD. Each word is assigned one or more sets of feature bundles consisting of several attribute-value pairs. For instance,

[6] http://www.iai-sb.de/iai/index.php/DUDEN-Korrektor.html, Haller et al., 2004.

```
{wnra=1,wnrr=1,snr=1,ori=Der,gra=cap,c=w,sc=rel,
»     ehead={g=m,nb=sg,case=nom};{g=f,nb=sg,case=gen;dat},
lu=der}
{wnra=1,wnrr=1,snr=1,ori=Der,gra=cap,c=w,sc=art,
»     ehead={g=f,nb=sg,case=dat;gen,infl=weak};
»     {g=m,nb=sg,case=nom,infl=weak};
»     {g=f;m;n,nb=plu,case=gen,infl=weak},
lu=der}

{wnra=2,wnrr=2,snr=1,ori=Weg,gra=cap,c=noun,nb=sg,case=nom;dat;acc,lu=weg}
{wnra=2,wnrr=2,snr=1,ori=Weg,gra=cap,c=vpref,lu=weg,error=44}
{wnra=2,wnrr=2,snr=1,ori=Weg,gra=cap,c=post,ehead={case=nil},lu=weg,error=44}

{wnra=3,wnrr=3,snr=1,ori=ist,gra=small,c=verb,vtyp=fiv,tns=pres,mode=ind,per=3,nb=sg,lu=sein}

{wnra=4,wnrr=4,snr=1,ori=frei,gra=small,c=adv,lu=frei}
{wnra=4,wnrr=4,snr=1,ori=frei,gra=small,c=vpref,lu=frei}
{wnra=4,wnrr=4,snr=1,ori=frei,gra=small,c=verb,vtyp=imperativ,nb=sg,lu=freien}
{wnra=4,wnrr=4,snr=1,ori=frei,gra=small,c=verb,vtyp=fiv,tns=pres,nb=sg,per=1,lu=freien}

{wnra=5,wnrr=5,snr=1,ori=.,gra=other,lu=.,c=w,sc=punct}
```

Figure 6.3: Analysis of the German sentence *Der Weg ist frei* by the MPRO module in the MPRO-KURD linguistic annotation solution.

the attributes *wnra* and *wnrr* stand for word number absolute and word number relative, and indicate respectively the position that the word occupies in the text and in the sentence. There are also ambiguous readings for the words *Der* [the, that, which] and the word *frei* [free, as a verb, as a particle and as an adjective/adverb].

(1) Der Weg ist frei.
    The way is free.

As for the linguistic features, there are attributes such as *ori*, the word as it appears in the text; *gra*, the presentation of the word in terms of case (capitalised, upper, lower, etc.); *c*, the category of the word – noun, verb, adverb (adv), etc.; *sc*, stands for subcategory, and it includes features as *rel*, *art*, etc. The feature *lu*, lexical unit, contains the lemma of the word.

There are certain features that depend on the category of the word. For instance, nouns, adjectives and determiners present the attribute *ehead*, that contains the typical agreement-related features such as number, gender and case, while verbs present features such as *vtyp*, verb type, with values such as fiv (finite verb), imperative, etc.

The output of the MPRO modules is passed on to KURD-based modules, which perform further disambiguation operations, and context-sensitive spell checking. Figure 6.4 presents the result of applying the KURD disambiguation module to the example sentence, where some readings were discarded and only the most plausible given the context are kept. For instance, the words *Der Weg* [the way] present only their nominative masculine singular readings. This decision is possible because they are followed by the finite verb form *ist* [is] of the copulative verb *sein* [to be] and *frei* [free], a word that has an adverb or verbal prefix reading.

```
3  {ori=Der,lu=der,c=w,sc=art,spec=def,ehead={nb=sg,infl=weak,case=nom,g=m},gra=cap}
4  ,{ori=Weg,lu=weg,c=noun,ehead={nb=sg,infl=weak,case=nom,g=m}}
5  ,{ori=ist,lu=sein,c=verb,sc=verb,vtyp=fiv,nb=sg,per=3,tns=pres,mode=ind,gra=small}
6  ,{ori=frei,lu=frei,c=adv,gra=small
7          ;{ori=frei,lu=frei,c=vpref,pref=vzs,gra=small}
8  ,{ori=.,lu=.,c=w,sc=punct,gra=other}
```

Figure 6.4: Analysis of the German sentence *Der Weg ist frei* by the KURD-based Morphological Disambiguator.

```
3  Disambiguate Noun Phrase =
4        Ae {c=w,sc=art,agr=_AGR},
5        Ae {c=noun,agr=_AGR},
6        a {c=verb,vtyp=fiv,lu=sein;bleiben},
7        a {c=adv;vpref}
8        :
9        Au {agr=_AGR}.
```

Figure 6.5: KURD rule to disambiguate the readings of the words *Der Weg* to their nominative masculine singular readings in the sentence *Der Weg ist frei.*

6.3.2.1.1 The KURD rule formalism

Since KURD is the formalism that we use to process texts beyond morphological analysis within a particular domain in the strategies exemplified in Chapter 8, we introduce the KURD formalism and explain which is the rule that applies in such a context. Technically, KURD is implemented as a set of finite-state machines that are sequentially applied. The grammar writer decides the order in which the rules are applied. The system applies a particular grammar on a text as far as there are words – the basic object of analysis – whose information is modified. After two continuous iterations with no modifications the algorithm stops the process (Carl and Schmidt-Wigger, 1998).

Figure 6.5 shows a sample KURD rule. KURD rules consist of three parts, as show in (6.1). Each rule has a NAME to identify it, a DESCRIPTION part and an ACTION part.

$$Name = Condition : Action. \tag{6.1}$$

The DESCRIPTION consists of a number of conditions that must match successive word descriptors (the feature bundles associated with each word). During the matching, the word descriptors that match are marked in order to be able to perform operations on them in the action part. A rule fails if a condition of the description part does not match, and then the action part does not apply.

In Figure 6.5 we observe how in the DESCRIPTION part (lines 4 to 7, the first column contains file line numbers) requires that for the rule to apply there is a sequence of two word descriptors consisting of a determiner (c=w,sc=art) and a noun with compatible agreement features – which is indicated by a binding variable (_AGR). If this is the case, then these two word descriptors are marked with an A (markers in KURD are capital letters from A to Z). The third and fourth condition require that there is a sequence of a finite verb form of the verbs *sein* [be] or *bleiben* [remain]. A crucial distinction between the first and second condition and the third and fourth is that for the first two an existential quantifier is used (e), while for the last two a wholistic quantifier is used *a*. Thus the rule requires that the readings of the third and fourth word descriptors are unambiguous for it to apply.

As for the ACTION part it simply contains a line that executes the *unification* operation on the variable _AGR. The unification operation is an essential operation in computation and particularly in computational linguistics that requires that terms to be compared be compatible. For instance, if applied to the values of the attribute gender two word descriptors unify if they have the same or compatible gender attribute values – for instance, masculine singular is compatible with masculine singular. When unification is applied to a variable, the algorithm checks that the values of all the relevant attributes are compatible. Those values that are not compatible are discarded and only those that unify are kept. There are many other operations available in KURD – the ones that give the name to formalism are Kill, Unify, Replace and Delete, but there are many others such as append, insert, generate (a mother), etc.Further details on MPRO and KURD can be found in (Maas, 1996; Carl and Schmidt-Wigger, 1998; Garnier et al., 2003a,b).

6.3.2.2 The CG-based solution

The CG-based solution is a piece of software that consists of several modules implemented using general programming languages and the formalism known as the Constraint Grammar – Karlsson (1990); Karlsson et al. (1995). The CG-based solution is used to analyse and spell check Catalan unrestricted text (Badia et al., 2001; Alsina et al., 2002; Badia et al., 2004).

This solution includes all the modules shown in Figure 8.5. The Tokeniser, the Morphological Analyser and the Spell Checker are implemented in Perl and C++ – including an algorithm based on minimum edit distance measures for the generation of correction proposals (Badia et al., 2004). The dictionary look-up process uses a word-form list that has more than one million entries and was generated with a two-way morphological processing module (Badia et al., 1997). CG-based grammars are used in the Morphological Disambiguator, the Context-Sensitive Spell Checker, and the Information Extraction module.

In contrast to the MPRO-KURD solution the CG-based one does not use explicit attribute-value pairs. It shows only the values – attributes are implicit –, but as we said in footnote [4] (p. 94) data representation and data structure are not necessarily related. Figure 6.6 shows the results of Tokeniser and Morphological Analyser for the sentence in (2).

(2) La casa és verda.
The house is green.

Linguistic information is added in any preferred systematic order, except for the fact that lemmata have to be at the beginning of each of the readings of the word, as shown in Figure 6.6 in the indented lines. For instance the word *La* has two readings: a determiner reading and a pronoun reading, both feminine singular. In the CG terminology a word with its associated readings is called *cohort*. For instance, *casa* and its three readings form a cohort.

```
<p id="1">
<s id="1">
"<La>"
                "el" Det fem sg
                "lo" Pron person febl acus 3pers fem sg
"<casa>"
                "casa" Nom com fem sg N5-FS
                "casar" Verb MInd Pres 3pers sg
                "casar" Verb MImp Pres 2pers sg
"<és>"
                "ser" Verb MInd Pres 3pers sg
"<verda>"
                "verd" Adj qual fem sg
" <$.>"
< /p>
< /s>
```

Figure 6.6: Results of the tokenisation and morphological analysis process for the Catalan sentence *La casa és verda*.

At this point of the processing, the modules implemented in Constraint Grammar are used. Karlsson et al. (1995: p. 1) define Constraint Grammar as "a language-independent formalism for surface-oriented, morphology-based parsing of unrestricted text. [...] All relevant structure is assigned via [...] simple mappings from morphology to syntax. The constraints discard as many alternative readings as possible [...] with the proviso that no genuine ambiguities should be obliterated".

As with KURD, what is crucial in this definition is that CG relies initially on morphological information to perform increasingly complex levels of automated analysis. As shown in the most recent versions of some of the products offered by the company that distributes a commercial licence of CG, Connexor Oy[7], CG-based grammars can be used for tasks as complex as functional dependency parsing, or semantic role labelling. With such techniques Connexor Oy can provide solutions for the identification of opinions (in several types of texts), detection of fraud, or extraction of specific knowledge from large collections of biomedical articles.

6.3.2.2.1 The CG rule formalism

Technically, the CG formalism is implemented as a set of finite-state cascades that are sequentially applied. The grammar writer does not decide the order in which the rules are applied. However, the grammar writer can decide to group rules into blocks so that they apply in a given order. The CG interpreter builds up a cascade

[7] http://www.connexor.eu/technology/machinese/

of finite-state automata that is actually responsible for controlling the accepted or active paths – sequences of states given and input. The system applies a particular grammar on a text as far as there are words whose information is modified. After two continuous iterations with no modifications the algorithm stops the process.

The basic structure of CG rules is reflected in (3). The TARGET characterises the specific linguistic features that have to be met by the linguistic object on which the action of the rule will be applied. The OPERATOR indicates which is the action to be performed on the TARGET in case the context matches. Possible actions are REMOVE, SELECT – for disambiguation –, ADD, MAP or REPLACE – information mapping. The CONTEXT defines the linguistic properties of the words surrounding the TARGET that need to be matched for the rule to apply.

(3) OPERATOR (TARGET) IF CONTEXT;

CONTEXT positions are indicated with positive (right of target) or negative (left of target) integers. The CG formalism provides the grammar writer with other functionalities, such as the possibility to work with relative or absolute positions, or to create contexts in which one or more of the conditions of application can be defined within a range of positions. There is a functionality called "careful mode" that allows grammar writers to restrict application conditions, so that rules only apply if the condition matches unambiguously.

The rules that would be needed to disambiguate the words *La casa* in our sample sentence (2) are reflected in Figures 6.7 and 6.8. In Figure 6.7 we have a rule that removes the Pron(oun) reading of any cohort whose context complies with the following conditions: it has a feminine singular determiner reading, it has a sentence start one position to the left (-1), it has a cohort with a feminine singular noun reading one position to the right, and a non-ambiguous finite verb (the C in 2C stands for careful mode, see above).

```
REMOVE (Pron) IF (-1 SentenceStart) (0 DET + FS) (1 NOM + FS) (2C VFIN);
```

Figure 6.7: Disambiguation rule that applies to the word *La* to remove the pronoun reading in the analysis of the sentence *La casa és verda.*

Figure 6.8 is the rule that applies to the word *casa* so that its noun reading is selected, which has the consequence that its two verb readings are discarded. In this rule the description context uses also the careful mode (2C) and looks at positions at the right and the left-hand sides of the target word.

```
SELECT (Nom) IF (-2 SentenceStart) (-1C DET + FS) (0 NOM + FS) (2C VFIN);
```

Figure 6.8: Disambiguation rule applying to the word *casa* to select the noun reading in the analysis of the sentence *La casa és verda.*

Further details on the CG-based solution used in ALLES can be found in Badia et al. (2001), Alsina et al. (2002), and Badia et al. (2004).

6.3.3 KURD and CG for shallow semantic processing

In this section we show how both KURD and CG can be used to analyse response chunks to analyse responses with a focus on activity-specific linguistic structures. This kind of task in NLP is often called shallow semantic analysis, and this is what motivates the title of this section. However, we believe that semantic analysis is a task that (i) implies a much more complex task than what we present in Chapter 8, and (ii) it can lead non-NLP experts to expectations that do not match with the real capabilities of NLP tools. Because of this we will tend to call it domain-specific information extraction or activity-specific learner response assessment.

6.3.3.1 CG-based shallow semantic processing

As for the task of annotating beyond the morphosyntactic level, CG can easily add new levels of information to one or each of the readings in a cohort. This is done by creating a rule file that contains rules that apply to the text to be analysed as a whole and not sentence-wise, as is usually done. Then, using the ADD operator, one can process the analysed text in order to check for the presence or the absence of the relevant linguistic structures.

Figure 6.9, exemplifies four rules that were included in one of the Information Extraction modules for the analysis of the Catalan version of an ICALL activity that is described later on Section 7.2.2.3. The rules correspond to a part of the response where the learner is expected to end an email with a complimentary close. As shown in (4), the rules envisage four different ways of expressing that in Catalan in order to comply with the activity's requirements – all of which correspond more or less to the English *Yours sincerely,* or *Yours faithfully,*.

(4) a. Atentament,
    b. Cordialment,
    c. Ben cordialment,
    d. Salutacions,
    e. Salutacions cordials,

```
ADD (@:ComplClose) TARGET (Adv) IF
        (0 ATENTAMENT OR CORDIALMENT) (1 COMMA);
ADD (@:ComplClose) TARGET (Adv) IF
        (-1 BEN) (0 CORDIALMENT) (1 COMMA);
ADD (@:ComplClose) TARGET (Nom) IF
        (0 Nom + SALUTACIONS) (1 COMMA);
ADD (@:ComplClose) TARGET (Nom) IF
        (0 Nom + SALUTACIONS) (1 CORDIALS) (COMMA);
```

Figure 6.9: CG rules for the analysis of the complimentary close in a formal letter in Catalan.

After this set of rules and other similar rules are applied to detect the relevant parts of the response, another CG-based module using a different set of rules checks for the global response correctness. This will be described in Chapter 8, where we describe the pedagogically oriented design and implementation of an NLP-based feedback generation module.

6.3.3.2 KURD-based shallow semantic processing

The Information Extraction module is implemented in a slightly different way in KURD. As described in Boullosa, Quixal, Schmidt, Esteban, and Gil (2005: pp. 32–34), the KURD formalism was enhanced during the ALLES project with a so-called "discourse" module. With its discourse module, KURD is capable of generating analysis nodes, e.g., feature bundles, at the sentence level – instead of associating them with word readings.

We will show how the rule for analysing part of the sentence in (5) would be implemented. This sentence is one of the possible responses to an activity in which learners are required to produce a satisfaction questionnaire – the activity is presented and worked out later on in Chapters 7 and 8.

(5) a. How satisfied are you with Stanley Broadband?

As shown in line 2 of Figure 6.10, the rule name is *CustomerSatisf*. This rule checks for the presence of the expected words that refer to the satisfaction of the customer in the response. The rule is fairly simple. It checks for the sequence of words *satisfied are you with*, and it maps the code *CustSatisf* to all of them – line 9. In addition, it maps this information to the sentence node, identified by a special symbol ($-1) in line 10. The rule *CustomerSatisf* tells the algorithm to go on with the processing of the block of rules corresponding to the that part of the response in which *Prod*uct is referred to – which we do not show.

```
2  CustomerSatisf =
3         Ae{ori=How},
4         Ae{ori=satisfied},
5         Ae{ori=are},
6         Ae{ori=you}
7       :
8         $-1g{disc=CustSatisf},
9         Ar{disc=CustSatisf},
10        j(rule=@Prod).
```

Figure 6.10: KURD rules to process a part of a possible response to one of the ICALL activities later on presented and worked out in Chapters 7 and 8.

After the rules in the Information Extraction module are applied to process the sentence in (5) a set of response chunks are identified and the sentence can be passed on to the module that will check for the correctness of the response. The completely analysed version of the sentence is reflected in Figure 6.11. We see particularly in line 2 the attribute *RespOrder* that contains all the corresponding response elements.

In each of the other lines corresponding to analysed tokens – lines 3 to 10 – we can see that each of them is identified as a member of a response element in the *disc* attribute. The elements in the attribute disc are part of the connection between the linguistic analysis and the pedagogical objectives of the activities to be matched with linguistic-based information. The methodology that we propose to design and implement such rules is explained in Chapters 7 and 8.

```
2  {wnrr=-1,RespOrder=a_HowDegOne_CustSatisf_Product_HowDegTwo}
3  ,{ori=How,lu=how,c=det,sc=interr,disc=HowDegOne,(...)}
4  ,{ori=satisfied,lu=satisfy,c=verb,vtyp=ptc2,disc=CustSatisf,(...)}
5  ,{ori=are,lu=be,c=verb,vtyp=fiv,disc=CustSatisf,(...)}
6  ,{ori=you,lu=you,c=det,sc=pers,disc=CustSatisf,(...)}
7  ,{ori=with,lu=with,c=w,sc=p,disc=CustSatisf,(...)}
8  ,{ori=Stanley,c=noun,nb=sg,disc=Product,(...)}}
9  ,{ori=Broadband,c=noun,nb=sg,disc=Product,(...)}}
10 ,{ori=?,lu=?,c=w,sc=punct,disc=HowDegTwo,(...)}
11 .
```

Figure 6.11: Linguistic analysis for the sentence (5) including the response elements detected by the Information Extraction Module.

6.4 Chapter summary

In this chapter, we introduced ALLES, the research setting in which our methodological instruments for the design and implementation of ICALL materials are exemplified. This context arises from a multidisciplinary research project carried out by a team of experts in several domains, among them experts in FLTL and NLP. My role in the project was to design and develop pedagogically informed NLP strategies for the generation of automatic feedback. These strategies are based on surface shallow semantic processing techniques for the automatic evaluation of learner responses, and implement summative and formative assessment strategies.

We presented the pedagogical concept underlying the TBLT-driven materials that resulted from the initial design phase, and characterised them in terms of Estaire and Zanón (1994)'s framework. This characterisation determines aspects of the topic, and the general linguistic and communicative goals of tasks. This approach requires the design of a learning sequence and the design of overall strategy of assessment procedures. However, it does not characterise the contents expected in learner responses, nor specific criteria for correctness for each response item. An approach supplying the instruments for a principled and formal characterisation of these latter aspects is the purpose of our research in the following chapters.

This chapter also introduced the NLP tools that serve as the basis for the implementation of practical assessment functionalities for the materials developed within the ALLES project. The general architecture for the linguistic processing of text using finite-state automata and a mal-rule approach was instantiated in two different software solutions, used for different languages in the project. In Chapter 8 the different levels of information generated by such an architecture are strategically combined to respond to FLTL needs and assessment requirements.

Chapter 7

Designing ICALL tasks – Characterisation of pedagogical needs

This chapter introduces and exemplifies the frameworks that we propose to characterise tasks and learner responses during the design phase, as well as the relationships between them. This characterisation will be used to pedagogically motivate the requirements for the linguistic analysis and feedback generation modules of an ILTS.

The Task Analysis Framework (TAF) characterises activities from a general pedagogical and linguistic perspective in terms of learning goals, learning processes, and type of response required from the learner. The TAF serves two purposes: (i) to determine the degree of communicativeness of the FL learning activity; and (ii) to distinguish FL activities that are good candidates to being turned into ICALL activities, and those that are not – mainly due to the expected outcome. The TAF is exemplified in the analysis of a set of learning materials.

The Response Interpretation Framework (RIF) characterises expected learner responses and their assessment criteria in detail. By applying the RIF to a particular task, a set of objective criteria for correctness can be produced, and a set of learner responses can be anticipated. To exemplify its use, we apply the RIF to four activities that are representative of the different kinds of activities that might be considered for NLP-based automatic assessment.

7.1 TAF: Task Analysis Framework

7.1.1 Definition

The TAF characterises FL learning activities with (i) the goal to know whether they are communicative or non-communicative activities and (ii) the goal to select them as candidates for being corrected with NLP-based automatic assessment strategies.

The TAF consists of eight rubrics that result from a selection of features taken from the works by Ellis (2003: pp.8–21), Littlewood (2004), and Bachman and Palmer (1996: Ch. 3), described in Chapter 4. The eight rubrics are:

- **Description**: General, informal description of the FL learning task as to understand its goal and definition features.[1]

- **Focus**: Pedagogical objective of the task: language as a system, focus on form, language as a means of communication, focus on meaning, or both (Estaire and Zanón, 1994, Ellis, 2003: pp. 9–10, Littlewood, 2004).

- **Outcome**: Result or product to be obtained by the learner by completing the activity (Ellis, 2003: p. 10).

- **Processes**: Abilities, strategies and real-world processes, as labelled by Bachman and Palmer, that learners are expected to deploy to complete the activity (Ellis, 2003: p. 10, Bachman and Palmer, 1996: pp. 75–76).

- **Input**: The materials, the instructions, and/or the information that learners are given to complete an activity (Ellis, 2003: pp. 9–11 and 289–291, Bachman and Palmer, 1996: pp. 52–53).

- **Response type**: Responses might be selected from a set of given choices, constructed (limited or extended production responses), or intangible (Bachman and Palmer, 1996: pp. 53–54).

- **Teaching goal**: Following Littlewood (2004: p. 322), activities are classified in categories according to how they relate to the goal of language teaching: non-communicative learning, pre-communicative practice, communicative language practice, structured communication and authentic communication. We add the class "instructions" to refer to those parts of the learning materials used to guide learners through the task.

- **Assessment**: Formative or summative; individual, collective or cooperative; external or self-assessment.

The rubrics *Focus*, (communicative) *Outcome*, and *Processes* contribute most to the characterisation of the FL learning units from a perspective of the pedagogical approach. They provide a sense of the "taskness" inherent to each activity. The rubric *Teaching goal* contributes to a more pedagogical-methodological side of the characterisation, but provides a class rather than pedagogical features. Altogether they reflect whether a particular learning activity qualifies as a communicative FL learning task and in what terms.

The rubrics *Input* and *Response type* provide information on the contents and the language that learners are expected to process, as well as on the contents and the language that they are expected to produce. The nature of the information to be processed and the length of the responses are features that might be used as an initial filter to select activities suitable for NLP-based assessment.

[1]This rubric is the only one that is not explicitly included in any of the mentioned works, but it helps the ICALL developer to access a quick, simple description of the task's goals.

7.1.2 Applying the TAF to *Education and Training*

To exemplify the application of the TAF, we apply it to a set of learning materials developed in the ALLES project. In particular, we apply it to the the activities of the unit *Education and Training* for the learning of English, the one for which the language-independent specifications were described in Section 6.2.

As described in Sections 6.2.2 through 6.2.5, the learning unit *Education and Training* consists of four Subtasks and a Final Task. The presentation follows the ALLES activity naming conventions and structure (Subtask and Activity, see Section 6.2.5.1).

7.1.2.1 Introduction and pre-test

Table 7.1 shows the TAF analysis of the two activities in Subtask 0, an introduction to the contents of the unit and a language pretest. The former introduces the learning activities that will be required from the learner. It describes the nature of the tasks, the associated learning processes that the learner is expected to go through and the outcomes that he or she is expected to produce to complete the unit.

Subtask 0: Introduction and pre-test	
Act. 1	
	Description Unit workplan: presentation to the learner of the final task and the sequence of preparatory tasks
	Focus Meaning
	Outcome None in particular, several in each task
	Processes Understanding pedagogical instructions
	Input Text describing the unit's contents from the learner perspective
	Response type None
	Teaching goal Instructions
	Assessment None
Act. 2	
	Description Pre-test on a subset of the linguistic and thematic objectives of the learning unit
	Focus Form (system-referenced test)
	Outcome None
	Processes Use of the infinitive and gerund in a text describing someone's professional career
	Input Text with gaps and lexical stems provided in parenthesis
	Response type Limited production (fill in the blanks with one or two words)
	Teaching goal Non-communicative learning
	Assessment Summative

Table 7.1: Subtask 0 of the learning unit *Education and Training.*

Activity 1 in Subtask 0 describes to the learner a larger sequence of tasks whose aim is for her/him to practice several real-world and pedagogical processes. It is a preparation for a communicative task composed by several other pedagogical subtasks.

The second activity is a pre-test that is system-referenced and summative. This is a medium-stakes assessment activity with the sole goal to test learner competences in some very restricted linguistic items that are connected to the grammar and vocabulary goals of the unit, which was devised as means to measure learning gain.

7.1.2.2 "Having a well-motivated workforce"

Table 7.2 presents the activities corresponding to Subtask 1, entitled *Having a well-motivated workforce*. Subtask 1 includes seven activities, some focusing on meaning, and others on form.

As shown in the table, Subtask 1 includes activities such as Activities 1, 2, 5 and 6, where the learner is exposed to texts related to the topic of the unit: motivation of workforces or career profiles. These activities correlate with real world processes such as being able to understand the contents of a document that is relevant for *your* work. These are all activities in which either no response from the learner is required, as in Activity 1, or in which learner responses are restricted selection responses such as multiple choice or true/false exercises.

Subtask 1: Having a well-motivated workforce	
Act. 1	
	Description Reading and reflection on the reasons that keep employees motivated to keep their jobs
	Focus Meaning and topic
	Outcome None
	Processes Understanding a text on corporate management and employee satisfaction
	Input Text and instructions to promote reflection on the topic
	Response type None required
	Teaching goal Communicative language practice
	Assessment Formative

Table 7.2: Subtask 1 of the learning unit *Education and Training* (continues).

Act. 2	**Description** Reading comprehension of the text (Act. 1) **Focus** Meaning **Outcome** None **Processes** Understanding the main ideas of the text **Input** Sentences stating (true/false) facts that can be drawn from the text **Response type** Selected response (true or false) **Teaching goal** Pre-communicative practice **Assessment** Formative
Act. 3–4	**Description** Vocabulary exercise based on the text in Act. 1; matching words with definitions **Focus** Form **Outcome** None **Processes** Use the context to infer the meaning of a word **Input** A text with selected words highlighted in it **Response type** Selected response (drag and drop) **Teaching goal** Non-communicative learning **Assessment** Formative
Act. 5	**Description** Exposure to a set of employee files of the human resources dept. The learner is expected to recognise them as a text type (e.g., different from curricula). **Focus** Form (text genre) **Outcome** None **Processes** Identifying a set of texts as being of the same type and the text type that they are **Input** A set of texts describing people's career profiles and a question **Response type** Selected response (multiple choice) **Teaching goal** Non-communicative learning **Assessment** Formative

Table 7.2: Subtask 1 of the learning unit *Education and Training* (continues).

Act. 6	**Description** Assigning courses to employees according to their schedules **Focus** Meaning **Outcome** None **Processes** Understanding time expressions and being able to relate course timetables with employees' schedules **Input** Texts describing people's schedules, and texts describing course timetables **Response type** Selected response (drag and drop) **Teaching goal** Pre-communicative practice **Assessment** Formative
Act. 7	**Description** Workplan: presentation to the learner of the following preparatory task, Subtask 2 **Focus** Form and meaning **Outcome** None **Processes** Understanding pedagogical instructions **Input** Text describing the contents of the following task from the learner perspective **Response type** None **Teaching goal** Instructions **Assessment** None

Table 7.2: Subtask 1 of the learning unit *Education and Training.*

As shown in Table 7.2, Subtask 1 includes activities with strictly linguistic goals, as opposed to communicative goals, such as the vocabulary practices in Activities 3 and 4. These activities correlate with linguistic cognitive processes such as word sense and text type identification. Both activities require selection responses.

The last activity in Subtask 1 is Activity 7, which presents the workplan for the following task, and has a role similar to Activity 1 in Subtask 0. This kind of activity is used as a companion for the learner through the unit.

In terms of outcome, none of the activities in Subtask 1 has a communicative outcome, which does not prevent them from focus on communicative aspects of language, as the comprehension tasks in Activities 1, 2, 5 and 6. In all of them, the input for the learner is textual (to be read) and other cognitive non-linguistic skills are required.

7.1.2.3 "Recommend a course and ask for information"

Table 7.3 presents the five Activities in Subtask 2, which is entitled *Recommend a course and ask for information.* Some of the activities in Subtask 2 focus on meaning, and some on form.

Activities 1 and 2 in Subtask 2 correlate with real-world processes such as being able to identify the topic of a conversation relevant for one's work, or to understand its contents. In contrast, Activities 3 and 4 involve linguistic processes such as the recognition of formulaic expressions to be used in suggestions and recommendations. Activity 5 presents the workplan for the following subtask.

The activities in Subtask 2 do not require any productive skills, are all responded with selection responses and do not involve the production of an outcome. The input data that the learner is given is both aural and textual.

Subtask 2: Recommend a course and ask for information	
Act. 1	
	Description Identification of the topic of a conversation where a training course is being recommended
	Focus Meaning
	Outcome None
	Processes Understanding recommendations and preferences in a conversation
	Input The audio file of the conversation and a question
	Response type Selected response (multiple choice)
	Teaching goal Pre-communicative practice
	Assessment Formative
Act. 2	
	Description Listening comprehension on the conversation heard in the Act. 1
	Focus Meaning
	Outcome None
	Processes Understanding recommendations, preferences and decisions, as well as reasons to make them
	Input The audio of the conversation and some questions
	Response type Selected response (multiple choice)
	Teaching goal Pre-communicative practice
	Assessment Formative

Table 7.3: Subtask 2 of the learning unit *Education and Training* (continues).

Act. 3		
	Description	Identify the expressions used for recommending in the conversation in Act 1
	Focus	Form (exponents of function)
	Outcome	None
	Processes	Understanding and identifying suggestions and recommendations
	Input	Transcript of the conversation and a set of questions
	Response type	Selected response (multiple choice with multiple correct answers)
	Teaching goal	Non-communicative practice
	Assessment	Formative
Act. 4		
	Description	Identify some more expressions used to make recommendations
	Focus	Form (exponents of function)
	Outcome	None
	Processes	Understanding and identifying suggestions and recommendations
	Input	Email where an employee justifies her decision to take a course and a set of questions
	Response type	Selected response (drag and drop)
	Teaching goal	Non-communicative practice
	Assessment	Formative
Act. 5		
	Description	Workplan: presentation to the learner of the following preparatory task, Subtask 3
	Focus	Form and meaning
	Outcome	None
	Processes	Understanding pedagogical instructions
	Input	Text describing next task's contents from the learner perspective
	Response type	None
	Teaching goal	Instructions
	Assessment	None

Table 7.3: Subtask 2 of the learning unit *Education and Training*.

7.1.2.4 "Asking information about a course"

Table 7.4 presents the application of the TAF to the activities of Subtask 3, under the heading *Asking information about a course*.

Activities 1 and 2, focusing on form, are related with linguistic processes such as the use of formulaic expressions and vocabulary related to courses and course registration procedures. Activities 3 and 4 focus both on form and meaning and are oriented to prepare an oral conversation. As for Activities 5 and 6, they try to emulate a phone conversation where the learner asks for information about a course, though there is no actual dialogues, and the partner's answers are recorded. Activity 7 presents the workplan of the following task, the unit's Final Task.

The input the learner is given is basically aural, though some textual input is included. This subtask requires learners to produce oral and written language, though Activities 5 and 6 are the only ones in which a communicative outcome is produced. As for the response type, Subtask 3 includes activities requiring selected responses (only Activity 1), and limited production responses (all other activities).

SUBTASK 3: ASKING INFORMATION ABOUT A COURSE	
ACT. 1	**Description** In a recorded conversation identification of a set of linguistic structures to ask information about courses **Focus** Form (exponents of function) **Outcome** None **Processes** Understanding and identifying ways of asking for or giving information on training courses **Input** A recorded conversation and a set of unordered expressions and identification labels **Response type** Selected response (drag and drop) **Teaching goal** Pre-communicative practice **Assessment** Formative
ACT. 2	**Description** Vocabulary practice: asking for course information and registration procedures **Focus** Form **Outcome** None **Processes** Use of topic-specific vocabulary **Input** Transcription of the conversation heard in ACT. 1. Blanks to be filled in with words provided in a list **Response type** Limited production (fill in the blank with one to three words) **Teaching goal** Non-communicative practice. **Assessment** Formative

Table 7.4: Subtask 3 of the learning unit *Education and Training* (continues).

Act. 3–4	
	Description Practice oral expressions to ask for information about a course **Focus** Form and meaning **Outcome** Isolated communicative acts **Processes** Preparation of a speaking activity (next two activities) to ask information about a course **Input** Hints on what has to be asked **Response type** Limited production (one sentence) **Teaching goal** Pre-communicative practice **Assessment** Formative.
Act. 5–6	
	Description Role play activity to ask for information about a course on the phone **Focus** Meaning **Outcome** Speaker turns in a conversation **Processes** Speaking activity to ask information about a course (assuming speech recognition is available) **Input** Recorded utterances of the other participant in the conversation and hints on what has to be asked. **Response type** Limited production (one sentence) **Teaching goal** Communicative language practice **Assessment** Formative
Act. 7	
	Description Workplan: presentation to the learner of the following preparatory task, the Final Task **Focus** Meaning **Outcome** None **Processes** Understanding pedagogical instructions **Input** Text describing next task's contents from the learner perspective **Response type** None **Teaching goal** Instructions **Assessment** None

Table 7.4: Subtask 3 of the learning unit *Education and Training.*

7.1.2.5 "Registering for a course"

Table 7.5 presents the application of the TAF to the activities of the Final Task, whose title is *Registering for a course.* Since this is a final task, according to ALLES pedagogical concept, it must focus on meaning.

As shown in Table 7.5, Activity 1–2 is related to real-world processes such as being able to understand recommendations from one's manager, being able to understand a calendar page, and being able to select the appropriate courses given time constraints and the manager's advice.

FINAL TASK: REGISTERING FOR A COURSE	
ACT. 1–2	**Description** Writing an email to register for a course given certain (fictive) conditions: one's own schedule, the list of available courses and a piece of advice from one's own manager **Focus** Meaning **Outcome** An email apply for registration in a course **Processes** Understand the advice from a superior; understand course descriptions; understand month schedules **Input** A calendar page, an email offering courses from the human resources department and a recorded message that your (fictive) boss left on your voice mailbox. Hints on the information and the text structure to be included in the email. **Response type** Extended production (a formal email) **Teaching goal** Structured communication **Assessment** Summative and formative
ACT. 3	**Description** Post-test on a subset of the linguistic and thematic objectives of the learning unit **Focus** Form (system-reference test) **Outcome** None **Processes** Use of the infinitive and gerund in a text describing someone's professional career **Input** Text with gaps and lexical stems provided in parenthesis **Response type** Limited production (fill in the blanks with one or two words) **Teaching goal** Non-communicative practice **Assessment** Summative

Table 7.5: Final Task of the learning unit *Education and Training.*

The expected outcome of Activity 1–2 is an email that has to be sent to the human resources department of the fictive company required for registration in a course. The input that the learner is given is visual, textual and aural. The activity's outcome is a text with a communicative goal. The response type is an extended production response.

The Final Task includes also Activity 3, a post-test of the system-referenced type and summative. This is a medium-stakes assessment activity with the sole goal of testing learner competences in some very restricted linguistic items by comparing the learner's punctuation to the value obtained by the learner in the unit's pre-test.

7.1.2.6 *Education and Training* as a whole

As a TBI-based instruction material, ALLES includes preparatory tasks focusing on form and meaning. For instance, Subtask 2 (Table 7.3) starts with an activity that focuses on meaning where learners are exposed to the language used to describe and recommend courses. Then it continues with two activities with a focus on form where learners are required to pay attention to the linguistic structures used to express recommendations and describe courses. In contrast, in Subtask 3 (Table 7.4), learners practise certain constructions to ask for specific course information and end up doing a role play activity in which they simulate a phone call.

ALLES often includes activities where the learner is exposed to topical knowledge. This is the case for Activities 1 and 5 in Subtask 1, where learners are exposed to a topic and to text genres relevant for the interest area.

As for the type of ability that learners are required to develop, the four traditional skills are found in this unit: reading, listening and writing are notably present in language learning activities – as well as through all the ALLES materials. Speaking is only present in one of the learning activities (and in the whole ALLES materials in another two or three). Such activities relied on the use of a speech recognition system and did not include any sort of human-machine interaction at the conversational level.

In the unit *Education and Training*, a number of real world processes are required on the learner side to complete certain tasks. Some of them are close(r) to real-life communication, such as being able to understand the contents of a document that is relevant for one's work, or being able to understand a voice mail from one's manager. Others are purely linguistic or pedagogical tasks such as vocabulary practice or identification of formulaic expressions.

Finally, as for the response types, some learning activities require responses to be selected from a given choice (e.g., multiple choice, true or false, matching through drag and drop), others imply the generation of limited production responses, and a third type of activity involves the creation of extended production responses (e.g., emails, as in the Final Task, or reports – in other learning units in ALLES). There are also activities where learners are not required to perform any perceptible action, such as Activity 1 in Subtask 1, but an introspective one.

In sum, *Education and Training* is as an example of the kind of unit of work found in ALLES. In general, ALLES materials present a mixture of activities with different foci, requiring different abilities and processes, and expecting a range of response types from learners. In our opinion, ALLES exemplifies a common programme in real-world instruction settings.

7.1.3 FL learning tasks as candidates to become ICALL tasks

As we exemplified with the ALLES materials, FLTL materials that are conceived as CALL materials include tasks that do not require learners to react with a language mediated response: These are activities where no perceptible outcome is required, or activities where a selected response is expected. If no perceptible outcome is required no feedback can be generated. If it is a selected response, then it does not pose any challenge in terms of NLP.

There are good reasons to require learner actions that are not language-mediated (Pujolà (2001: p. 83), Ellis (2003: p. 15), and Littlewood (2004)), and the pedagogical design and development process, as well as the techniques employed to generate feedback in such kind of activity is discussed in the literature (see, for example, Arneil and Holmes, 1999 and Pujolà, 2001, 2002). Our research, however, centres on activities that require written language-mediated responses on the learner side that can be linguistically processed by computers.

Similarly to the ICALL systems discussed in Chapter 2, ALLES presents a mixture of activities, some of which require learners to produce language-mediated responses. Some of them require the production of limited production responses, such as fill-in-the-blank with one or more words or phrases. Others require extended productions responses such as paraphrasing activities or open questions, or even short composition activities. These are the types of tasks that represent a challenge in terms of NLP-based automatic assessment.

7.2 RIF: the Response Interpretation Framework

In this section, we present and exemplify the application of the Response Interpretation Framework (RIF), our proposal to characterise expected responses and the corresponding assessment criteria for a given ICALL-to-be FL task. The RIF characterises the pedagogical and linguistic properties of FL learning tasks, so that response variation, a key to NLP tractability, can be anticipated formally. Response anticipation yields a design-based (as opposed to corpus-based) gold standard set of responses, and helps establishing a set of objective assessment criteria.

7.2.1 Definition

The Response Interpretation Framework is a blend of the work by Bachman and Palmer (1996: pp. 53–56) and Estaire and Zanón (1994: pp. 4, 30, 49 and 58) using an NLP filter lens when looking at them. It identifies characteristics of the learning activities that have a greater influence on the definition of the topical and linguistic knowledge – the contents and the language – that learners are expected to produce (Bachman and Palmer, 1996: Ch. 3). Even though Bachman and Palmer present their test task characteristics framework for the analysis of tests, as suggested by Ellis (2003: pp. 312), their work can also be used to analyse tasks which are not necessarily criterion tests.

The RIF consists of six features (some of the concepts have already been introduced in Chapter 4 but are repeated here for the sake of convenience):

- **Instructions**: This feature includes properties as the **language** in which the instructions are given, the **channel** in which they are presented, and the **specification of the procedures and tasks**. Instructions can thus be "lengthy or brief; with or without examples; provided one at a time, linked to particular parts of the test [...]" (Bachman and Palmer, 1996: pp. 50–51).

- **Input**: This feature refers to the material that learners are expected to process to complete the task in terms of format (aural or visual; a word, a phrase, a sentence, etc.; live or recorded; etc.) and in terms of language (whether it corresponds to language knowledge such as vocabulary, morphology, syntax, rhetorical or conversational organization, or topical knowledge such as personal, cultural or technical information in it). Following Douglas (2000), we furthermore distinguish between **input data** – strictly the materials or objects that learners have to process to undertake the task – and **prompt** – material used to set up a specific communicative situation with no separate input data or where the input data is not enough.

- **Expected response**: This feature includes the **format**: oral or written, lengthy or short. The **language** of the response is described in the same terms as it is in the input. A very relevant characteristic is the **relationship between the input and the response**, which we analyse in terms of **scope of relationship** (which relates to the amount of input data that has to be processed in order to complete the task), and in terms of **directness of relationship** (direct responses are strictly related to the input, while indirect responses rely more on presupposed knowledge).

- **Thematic content of the response**: This feature analyses the information to be included in the response. It is related to what Bachman and Palmer (1996: p. 54) refer to as "topical characteristics of the language of the expected response", or topical knowledge. We divide it in **entities** and **relations**, as two different types of information expected. These are very common terms in NLP – particularly in Information Extraction tasks –, drawn from the field of compositional semantics, which are used to refer to real-world entities and the kinds of relations that can be established among them.

- **Linguistic content of the response**: This feature analyses the linguistic properties of the expected responses in terms of text structure and rhetorical organization, functional contents, grammatical content and lexical content – it is inspired by Estaire and Zanón (1994: pp. 30 and 58).

- **Assessment criteria**: This feature relates to how the learner output, the product, is evaluated in pedagogical terms by the ICALL system. We divide it into two subfeatures: **criteria for correctness** and **scoring procedure** (Bachman and Palmer, 1996: p. 52). The former helps determine the correctness of the response, and the latter the steps involved in scoring the tasks.

The first two fields, instructions and input, inform about the procedures and the materials that content designers expect learners to use to complete the task. The next three fields, expected response, thematic content of the response, and linguistic content of the response, explore the degree of variation that can be predicted in terms of topical and linguistic knowledge. The last field, assessment criteria, helps to specify what the response has to comply with to be correct, and how it is scored.

7.2.2 Applying the RIF to four FL learning tasks

The four tasks to which the RIF is applied present different degrees of complexity, given their pedagogical characteristics, the properties of their expected answers, the type of assessment required. In addition, they all comply with the condition to be communicative tasks that require production responses, of at least one sentence.

7.2.2.1 Task type I

The task that exemplifies Task type I corresponds to Activity 4 of Subtask 3 in *Customer Satisfaction and International Communication* as found in the ALLES website and is entitled *Stanley Broadband customer satisfaction questionnaire*. The TAF analysis of this task, a step we consider previous to the application of the RIF, is the following:

Description Writing the questions that will be included in a customer satisfaction questionnaire to distribute among customers of the company that offers the Stanley Broadband service

Focus Meaning and form

Outcome A customer satisfaction questionnaire, by producing five interrogative sentences that will be part of it

Processes Ask people about their opinion and intention with respect to a product or service; use of relevant linguistic structures to ask for information

Input With a total of five items, the task presents a separate item for each response. For each of them a blank space is provide with a hint on what is it that the learner should ask about and on certain lexical material to be included in the response.

Response type Limited production response: a sentence per item

Teaching goal Communicative language practice

Assessment Formative

The TAF characterisation shows that this is a writing task, focusing on meaning and form, that it has a communicative outcome, and that its goals are related to real-life processes, particularly to real-life processes in the marketing branch. In terms of the responses, limited production responses are expected, and a total of five items are separately required to be answered, each of which includes some hints for the learner. From this latter information, we infer that the learner's work depends to a certain extent on the input (prompts, instructions or hints) provided. In terms of assessment, being it a preparatory task in ALLES, it requires formative assessment.

With this information, this task can be considered a candidate to become an ICALL task. Therefore, the RIF can be applied to further learn about its pedagogical goals and the linguistic needs.

7.2.2.1.1 Characterisation of the response in pedagogical and linguistic terms

Table 7.6 exemplifies the application of part of the RIF analysis to all the task items as a whole, with the aim of anticipating the nature of the expected response. The task contains a prompt, instructions, an example, input data and space for the response (identified in the column on the right). For the sake of clarity, the table restricts itself to the first three items of the RIF (instructions, input and expected response), while the other three (thematic content, linguistic content and assessment criteria) are analysed separately below.

The prompt identified in Table 7.6 requires the learner to place himself or herself in a setting where a customer satisfaction questionnaire has to be produced: A kind of role play activity to be performed individually. The language of the prompt, the instructions and the input data is English, and the channel is textual. As for the specification of the procedures and tasks, the instructions include an example that shows learners the production procedure expected from them.

The instructions of the task require learners to use a set of clues given for five different items to produce five interrogative sentences for the questionnaire. All items include input data that follow the pattern "Ask about X" and "Use the expression or word Y in your answer". Learners cannot decide on their own – e.g., on their previous experience in this area – the topics to be asked about; moreover, they are required to use certain words as part of their responses.

In terms of expected response, it exemplifies tasks with a narrow scope of relationship between input and response – in Bachman and Palmer's terms (1996: pp. 54–56). To complete each response item learners have to read the ask-about-X and the use-expression-Y instructions and then produce an interrogative sentence. The information to be processed is short, and the required response is also short.

As for the directness of the relationship, all responses in all items present a direct relationship to the input data in terms of topical and language knowledge. Topical knowledge is notably restricted in the input given to learners, and language knowledge is partially restricted by it. In general, there is not much room for creativity.

The following step in the RIF is to characterise the thematic and linguistic contents to be included in the response, from which a list of assessment criteria will emerge. For space reasons, this part of the RIF is only applied to one of the items. Table 7.7 presents schematically the application of the rest of the fields in the RIF to Item no. 1. The first block in Table 7.7 analyses the thematic content of the response. The second block analyses the linguistic content of the response.

Customer Service and International Communication,
Subtask 3, Activity 4

Prompt	Imagine you work for Stanley Broadband. You have just listened to the interviews with Trevor and Janet. You would like to improve the service that Stanley Broadband offers. You have to compose a questionnaire to find out more about how to improve your company's service.
Instructions	Your task is to use the clues given for each box and write the necessary question.
Example	0. Ask what customers thought about the cost of Stanley Broadband compared to other companies who provide Internet services. Include the words *Did you ... find ... expensive...*? *Did you find Stanley broadband more expensive than other broadband service providers?*
Input data	1. Ask about customers level of satisfaction with Stanley Broadband. Write a question beginning with *How....?*
Response	
Input data	2. Ask about customers favourite feature of the Stanley Broadband service. Include the words *What ... best*? in your question.
Response	
Input data	3. Ask what customers don't like about the Stanley Broadband service. Include the words *What ... least*? in your question.
Response	
Input data	4. Ask about frequency of Internet usage. Begin the question with the words *How often* ...?
Response	
Input data	5. Ask customers to describe future improvements they would like to see in the Stanley Broadband service. Begin the question with *What improvements* ...?
Response	

Table 7.6: Partial RIF characterisation of Activity 4 in Subtask 3 in *Customer Satisfaction and International Communication.*

Thematic content of the expected response	
Entities	– Stanley Broadband – Your interviewee (your customer) – The interviewer
Relations	– Interviewee has an opinion or an experience as user of Stanley Broadband – Interviewer asks about the the level of satisfaction of interviewee using Stanley Broadband

Linguistic content of the expected response	
Functional	– Ask customers about satisfaction with a product *How satisfied are you with ...* *How happy are you with ...* *How much did you like using...*
Syntactic	– Use word order of wh-questions – begin with *how*
Lexical	– *how, Stanley Broadband, you, satisfied, happy, satisfaction, ...*
Pragmatics	– Use the appropriate register. – Use an interrogative sentence beginning with *how*
Graphology	– Use the appropriate spelling.

Table 7.7: RIF analysis of Item 1 of Activity 4 in Subtask 3 of the unit *Customer Service and International Communication* in ALLES.

As shown in Table 7.7, the response to Item 1 has to include a reference to three different entities: *Stanley Broadband*, the product that the questionnaire is about; the *interviewer*, who might be absent linguistically but is or represents the entity requiring the information; and the *interviewee*, who coincides with the customer and is the person to whom the question is addressed.

As for the relations, there are two of them expected in the response. The first relation is related to the interviewee, as a user of the Stanley Broadband service, having an experience and an opinion about it. Therefore, the response requires a piece of language referring to an experience as a user of a service, such as *be satisfied with*, *be happy with*, *like*, or *enjoy*. The second relation shows the interest of the interviewer for the interviewee's degree of satisfaction: The learner has to include the inquiry of the customer's satisfaction.

As for the linguistic content of the response, from a functional point of view, the goal is to elicit from the learner a piece of language corresponding with the communicative function of asking someone (a customer) the opinion about a product or service. As shown in the table, asking about someone's opinion can be put into words using different exponents for the function such as *How satisfied are you with ...*, *How happy are you with ...*, or *How much did you like using...* Of course, this communicative function could be accomplished by constructions such as *What do you think of ...*, *What's your opinion about ...*, or *Are you satisfied with...?*, but the task instructions require the learner to use the wh-word *how*.

From a syntactic point of view, the response has to be a direct interrogative sentence, otherwise it could not – or at least hardly – start with *how*. That it must be an interrogative sentence can be drawn from the prompt – "you have to *compose a questionnaire*" –, the instructions – "write the *necessary question*" –, and the item's input data: – "*ask about* customers level of satisfaction". The response is expected to present the word order of direct interrogative sentences in English.

From a lexical point of view the response is expected to include the words *how, Stanley Broadband, you, and satisfied*, but it could also include variations such as *happy, satisfaction, the Stanley Broadband service,* etc. Lexical content will be an important aspect to take into account when anticipating learner responses.

Finally, the last two aspects analysed in terms of linguistic content in Table 7.7 are pragmatics and graphology. They derive from the communicative setting described in the activity's instructions and the prompt. The response has to have the form of a question because to elicit an open answer on the customer side is the most commonly accepted way to do it, at least in the Western world, addressing an English speaking audience. From the perspective of communicative appropriateness, one should also consider politeness and observance of socially relevant linguistic norms, and that is why graphology is important.

7.2.2.1.2 Assessment

With the previous analysis at hand, the assessment criteria can be specified. Because of ALLES specifications (see Section 6.2.6), this Activity requires formative assessment. Therefore, only the criteria for correctness will be specified.

To produce a correct answer a learner response should consist of:

- An interrogative sentence:
  - asking for the level of satisfaction of the askee,
  - including a reference to the Stanley Broadband service, and
  - starting with *how*.

In addition, given the simulated communicative setting, the following criteria should be taken into account:

- use of the appropriate word order in interrogative sentences,
- use of the appropriate register,
- use of the appropriate spelling.

7.2.2.1.3 A RIF-based list of predictable responses

With such an analysis, a set of correct responses to Item 1 can be generated. Two of the possible correct responses would be the ones in (6).

(6) a. How satisfied are you with the Stanley Broadband service?

b. How satisfied are you with Stanley Broadband?

Two alternative correct responses would be the ones presented in (7). In (7a) the word "satisfied" in (6) has been replaced with "happy", a synonym. In (7b) the structure used to express satisfaction has changed. It is based on the use of the verb "enjoy", and it results into "enjoy (using) X".

(7) a. How happy are you with Stanley Broadband?
b. How much did you enjoy using the Stanley Broadband service?

Obviously, the number of alternative correct responses is much larger than the one reflected here, but it is certainly and reasonably discrete. Moreover, after a certain period of using this same task with a reasonable number of learners, a number of alternatives will be observed as the most usual. In Chapter 9, we describe how learner responses can help enlarge both the set of gold-standard responses and the list of most probable errors or incorrect responses.

7.2.2.2 Task type II

The task that exemplifies Task type II corresponds to Activities 5 and 6 in Subtask 1 of the learning unit *Company Organization* in ALLES. The task's title is *Describe the structure of your company to a colleague of yours*. As we did before, we first present the TAF analysis:

Description The learner is expected to write an email to a colleague of hers (Raymond) explaining him the structure of the (fictive) company they work in.

Focus Meaning.

Outcome An email describing the structure of the company.

Processes Writing relatively formal emails (it is a professional context); describing the structure of a company, with its departments, its delegates and the interrelations.

Input The task provides the learner with the company's organisation chart and a space to write the email. It includes some expressions the learner can use to describe responsibilities and interrelations in a company.

Response type Extended production response: an email.

Teaching goal Between communicative language practice and structured communication.

Assessment Formative.

According to the TAF analysis, this is a writing task that focuses on meaning, that has both a communicative goal and a communicative outcome, and that implies processes that are comparable to those in real-life. In terms of the response, it requires an extended production response, an email, which is considerably restricted by the input. It is like describing a picture. In terms of assessment, it requires also formative assessment.

Compared to the type I task, the response to the task exemplifying type II tasks is longer and more complex: an email compared to separate one-sentence questions in a questionnaire. Moreover, Activity type I is more restrictive on form. In all other respects, they are very similar, since they both restrict considerably the thematic and linguistic contents of the responses, they both focus on meaning, and they both require formative assessment because it is part of a preparatory task.

7.2.2.2.1 Characterisation of the response in pedagogical and linguistic terms

Table 7.8 shows the application of the RIF to this task. The prompt of the activity requires the learner to send an email to a colleague of hers/his named Raymond, in which s/he is required to describe the structure of the company. It is again a kind of a role-play activity to be performed individually. The learner is put in a setting where someone needs to be introduced to the structure of the company.

The activity's instructions require the learner to make sure that certain pieces of information are provided. These are included in or inferable from the organization chart of the company, identified as input data in Table 7.8. The instructions also suggest learners use certain linguistic elements that have been introduced and practised in previous tasks (cf. Subtask 1 in *Company Organization* in http://www.iai-sb.de/alles).

In terms of the expected response, the activity presents a moderately broad scope. To complete it learners need to pay attention to the instructions and the input data, which are rather concise. On the basis of that, their texts can expand on the information provided, and the responses to be elaborated are longer and more complex texts, at least in terms of linguistic knowledge.

As for the directness of the relationship, the task presents a dual nature. In terms of thematic content, the information to be provided is directly related to the input data, mainly the chart. In contrast, the expected linguistic content is vaguely restricted in terms of the lexical contents to be used to describe the relations between entities such as the departments and the people in charge of them. Other linguistic aspects such as the email structure or specific ways of linking and wording the expected information are not made explicit and are left open. The learner can decide on her/his own on the linguistic resources to be employed.

The first block in Table 7.9 characterises the thematic content of the response. It shows the expected entities and the relations:

- A reference to the own company
- The number of departments in the company, their names and subdepartments
- The person responsible for each department

Company Organisation, Subtask 1, Activities 5 and 6

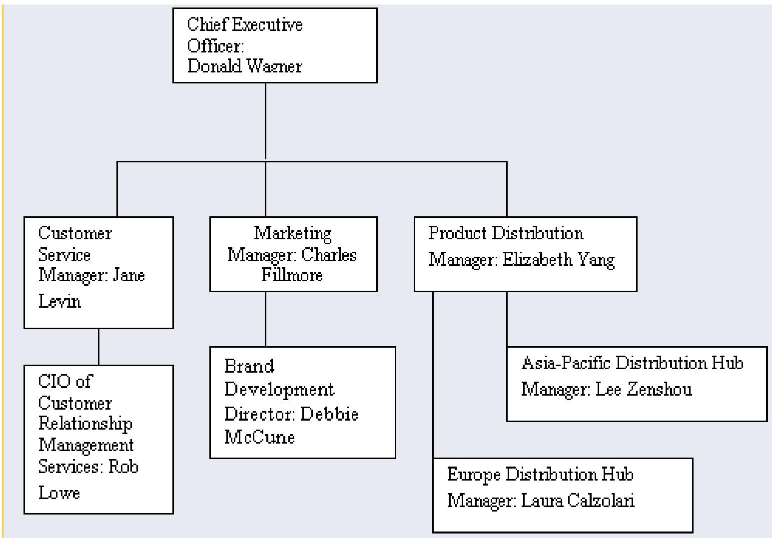
Prompt

Take a look at the chart below. The chart describes the structure of the company Jamdat Mobile. Pay attention to the number of departments and who reports to whom. Once you have carefully reviewed the chart, click on the arrow in the upper right corner to go to next screen and there you have to send an email to your colleague Raymond and describe the structure of Jamdat Mobile.

Instruction

Hint: Do not forget to describe how many departments this company has, who is reporting to whom etc. To describe the structure, you can use the expressions such as “to be in charge of”, “to report to”, “to be responsible for”, etc.

Input data



Response

From:

To:

Subject:

Table 7.8: Application of the RIF to Activities 5 and 6 in Subtask 1 in *Company Organisation* in ALLES.

- Responsibilities and accountability to third parties for each position

The names of the departments, the people and the relations among them are literally specified in the chart provided as input data. The degree of variation related to the number of entities and relations is delimited by the organisation chart.

The second block in Table 7.9 describes the linguistic content of the response. In terms of functional content, the activity aims to elicit from the learner linguistic expressions to describe a company, its departments and its names: *X is divided in Y departments*, *X has Y departments, namely, P, Q, R.*. It has to include the department heads and their interrelations: *Department X is led by Y, who is in charge of P and Q*. The corresponding exponents of function would be used to describe dependencies between departments and people, people's responsibilities, or the greeting and the closing sections in the email.

As for syntactic content, learners are expected to use the present simple to describe states of affairs, passive and active structures as part of descriptive texts, subordinates, coordinates and juxtapositions to express dependencies or department or peoples properties, and the functional words – prepositions, conjunctions – required by the verbs, nouns, adjectives or subordinates relevant for task goals.

Thematic content of the expected response	
Entities	– Company name: Jamdat Mobile, Jamdat. – Company's Chief Executive Officer (CEO): Donald Wagner. – Top-level departments: Customer Service, Marketing and Product Distribution – Second-level departments: Customer Relationship Management, Brand Development, Asia Pacific Distribution Hub and Europe Distribution Hub – Personnel: Jane Levin, Charles Fillmore, Elisabeth Yang, Debbie McCune, Rob Lowe, Lee Zenshou and Laura Calzolari. – The email's addressee: Raymond

Table 7.9: Application of the RIF analysis to Activities 5 and 6 in Subtask 1 in the unit *Company Organization* in ALLES (continues).

Relations	– The company has three departments – The company has a Chief Executive Officer – The Chief Executive Officer is Donald Wagner – Customer Service is led by Jane Levin – Marketing is led by Charles Fillmore – Product Distribution is led by Elisabeth Yang – Jane Levin, Charles Fillmore and Elisabeth Yang report to the CEO. – Brand Development is led by Debbie McCune – McCune reports to Charles Fillmore – Customer Relationship Management is led by Rob Lowe – Rob Lowe reports to Jane Levin – Asia Pacific Distribution Hub is led by Lee Zenshou – Europe Distribution Hub is led by Laura Calzolari – L. Zenshou and L. Calzolari report to Elisabeth Yang
Linguistic content of the expected response	
Functional	– Describe the structure of a company – Describe departments in a company *Our company has X departments.* – Name the departments in a company *The departments are: X, Y and Z.* – Name the managers of each department in a company *The department of X is lead by Y.* – Describe dependencies between departments/people *The department of X has X subdepartments: (...)* – Describe people's responsibilities in a company – Greeting, closing and signing in formal emails.
Syntactic	– Present simple – Passive and active structures – Use of subordinates – Prepositions governed by relevant verbs
Lexical	*– to be in charge of, to report to, to be responsible for, to coordinate, to have ... departments, to depend on, to delegate to, department, head, subdepartment, area, ...*
Pragmatics	– Description structure: either bottom-up or top-down, but coherently structured – Use adequate discourse markers and pronouns to glider the text – Email structure: greeting, body, complimentary close, signature

Table 7.9: Application of the RIF analysis to Activities 5 and 6 in Subtask 1 in the unit *Company Organization* in ALLES (continues).

GRAPHOLOGY	– Observe grammar and spelling required in private professional contexts

Table 7.9: Application of the RIF analysis to Activities 5 and 6 in Subtask 1 in the unit *Company Organization* in ALLES.

In terms of lexical content, learners are expected to produce expressions within the semantic field of companies. To express inclusion or composition relations they can use expressions such as *X has Y*, *X consists of Y ...*, *there are X ... in Y*. To express dependencies between departments or people they can use expressions such as *department X has Y subdepartments*, *department X depends on department Y*, *X delegates to Y part of his/her work*, *X reports to Y*, *Y is reported to by X*, *X coordinates Y*. In this respect, instructions encourage ("you can use ...") the learners to use expressions such as *to be in charge of*, *to report to*, or *to be responsible for*, but there is some room for creativity.

At the pragmatic level, learners are required to provide their (fictive) email address, the email address of their colleague (Raymond) and the subject of the email. These are all parts of an email in a standard professional communicative setting. As for the email itself, it should contain (i) a greeting, (ii) the body of the message containing the expected information, (iii) a complimentary close, and (iv) a signature.

In terms of information structure, also at the pragmatic level, one would expect coherence and cohesion in the description. As shown in the last block in Table 7.9, one possibility is to start the description at the lower level departments and end with the upper level departments. It would be unusual, probably unacceptable, to describe things in a mixed order.

Since the task emulates a professional context, even if the email is sent to a colleague, it has to comply with certain formal requirements. The context features allow us to predict that there should be sentences such us "Dear X", or "Dear Mr. X", as required in professional communication, even if private. Formulaic expressions such as "Sincerely yours" or "Sincerely" as a complimentary close, and a proper name in the signature, are also expected. The instructions do not specify it, but one could expect to find contact details after the signature, or even a legal note on the confidentiality of the information.

7.2.2.2.2 Assessment

The criteria for correctness for this task can now be specified. For a response to be correct it should include:

- An email containing addressee, subject and text
- The text email should contain:
  - Greeting, probably including *Raymond*

  - Body, including all the information reflected in the picture (number of departments, names of the heads, Chief Executive Officer in the company, etc.)
  - Complimentary close
  - Signature

- In terms of language knowledge the response has to include:
  - The appropriate expressions to describe relations between company departments, company colleagues, etc.
  - Use the appropriate syntactic structures in accordance with the lexical choices
  - Structure the text in a coherent and cohesive manner

7.2.2.2.3 A RIF-based list of predictable responses

With this information we can build a set of possible correct responses. However, for this response this task is much more difficult, given its length and the amount of thematic and linguistic content to be included. Even though three out of the four parts of the email (greeting, complimentary close and signature) are very restricted in form and content, the fourth one, the body of the email is fairly open. The body of the email is restricted in terms of thematic content, but it requires a large number of entities and relations to be expressed. The corresponding different ways of wording each of these entities and relations are relatively open.

To show that it can be done, however, we provide in (8) a sample response, which was produced by one of the content designers of the ALLES materials.

(8) Dear Raymond,
I am going to describe how the structure in Jamdat works. From top to bottom, the CEO is Donald Wagner, who coordinates three areas: Customer Service, Marketing, and Product Distribution. The person in charge of Customer Service is Jane Levin. She reports to Mr. Wagner about clients and she delegates to Rob Lowe, who is responsible for Customer Relationship Management. Charles Fillmore is the Marketing Manager. The department of Brand Development is managed by Debbie McCune, who reports to Mr. Fillmore. The head of Product Distribution is Elizabeth Yang, who is reported to by Lee Zenshou, for the Asian-Pacific area, and Laura Calzolari, for Europe. I hope this summary is clear for you.
Best regards,
SIGNATURE

In this sample response, we observe linguistic characteristics that are not reflected, or at least not explicitly, in the RIF analysis. For instance, there is a sentence introducing the topic of the text and the writer's goal at the same time: *I am going to describe how the structure in Jamdat works.* An introductory sentence as such is very common in letters and emails to introduce the topic of the message to the addressee – e.g., *I am dropping a line to tell you ...*, *In response to your email, I ...*, or *Following our conversation from this morning ...* As a consequence, a relation such as "the author expresses the reason for writing an email to the addressee" should, or could, be part of the RIF analysis in terms of the thematic content of the response.

In terms of text structure, the description of the company's structure is presented top-down, which is even made explicit in the response sample, and left-to-right, only implicit. It could certainly be presented in different orders, some of which might be correct, others not.

In terms of lexical content, the word *areas* is used to describe the higher rank departments, those that present a straightforward dependence from the CEO. In addition, the word *department* is used for the departments under any of those three. Finally, the word *area* is used to describe what in the chart is labelled as *hub* – e.g., *Asia Pacific Distribution Hub* – to describe the subdepartments in the Product Distribution department.

Still at the level of lexical content, the text in (8) presents lexical choices whose meaning is close to the choices in the activity's prompt – *be in charge of* or *be responsible for*. This is the case for *coordinates*, which is not in the task instructions, but suggested in Table 7.9. It is also the case for *is managed by*, or *the head of X is Y*. The sample response uses also verb forms such as *delegates*, and *is reported to by*, which present a meaning symmetrical to *report to* in terms of the expressed relation.

By analysing the sample response an expansion of the RIF definition is possible. This expansion mainly determines possible alternative correct responses.

7.2.2.3 Task type III

The task that exemplifies Task type III corresponds to Activities 1 and 2 of the Final Task in *Education and Training* in the English version of the ALLES materials. The task's title is *Registering for a course*. We first present the TAF analysis:

Description The learner has to write an email to the human Resources department of the company for which she works and register for a course, taking into account a piece of information given in the input.

Focus Meaning.

Outcome An email requiring to be registered in a course.

Processes Writing a formal email; registering one for a course; being able to argue for the suitability and appropriateness of the course to one's own interests and restrictions; understanding messages from one's boss; understanding course descriptions; understanding calendar pages.

Input The task includes three pieces of input data: a calendar page with one's fictive month schedule; a recorded message from the department manager; and a description of the available courses. The instructions include also information on what has to be included in the email.

Response type Extended production response: an email.

Teaching goal Structured communication.

Assessment Summative and formative.

According to the TAF, this is a writing task focusing on meaning that has both a communicative goal and a communicative outcome. It is again an individual role play activity in which the learner is put in a setting where s/he has to integrate information provided by third parties (manager's voice message and course information) and her/his own (personal calendar) to be able to require for the registration per email to an in-company department. All of these processes are compatible with a professional setting.

As for the response, it would be qualified as an extended production response, an email, whose thematic content is considerably restricted by the input data and the task instructions (see the RIF analysis below). Since it is a final task in ALLES, it requires both summative and formative assessment. Compared to the two previous task types this task type has a longer response than Task type I, and a more complex response in terms of processes than both Task types I and II.

7.2.2.3.1 Characterisation of the response in pedagogical and linguistic terms

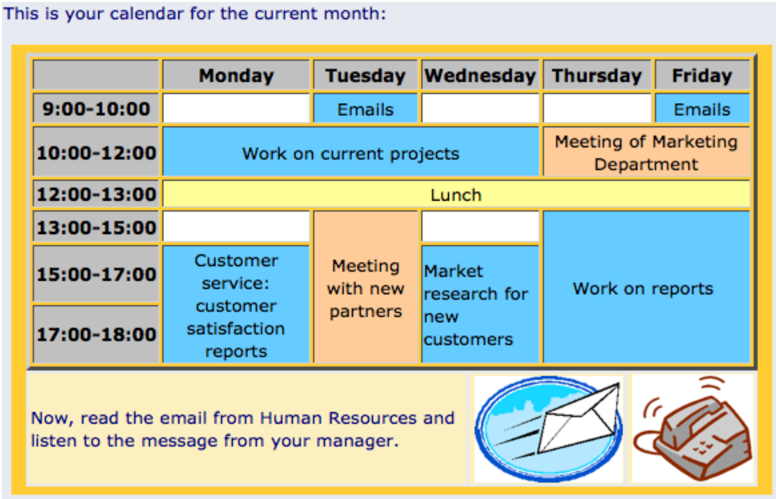
Table 7.10 shows the application of the RIF to characterise the expected response for Activities 1 and 2 in the Final Task of *Education and Training* in terms of instructions and input. The prompt requires the learner to send an email to the human resources department of her company. According to this, s/he has to express her interest for a course, explain that this course suits the advice given by her manager, as well as her monthly schedule. As seen in the image included in the *Input data*, to complete it the learner is provided with a calendar page, a recorded voice mail from her boss with information relevant for the decision, and a list of courses offered by the department (the two latter are reflected in the image in two icons in the lower right corner of the image: its actual contents can be found in Annex C).

As for the instructions, they require the learner to make sure that certain pieces of information are provided, most of which are included in or inferable from the information in the input data. Instructions also remind learners to include certain text elements in the email, which are mainly oriented to guide learners in providing the requirements of the text genre.

PROMPT	Now, you are an employee of the marketing department at Inteltrans. You just got an email from the Human Resources Department. In this email several courses for training and education of employees are listed. First, you need to check your calendar to see if you have some time free to take some courses. Then, you have to read the email from Human Resources and check for the schedule of the courses. Finally, you need to check your voice mail to listen to an important message from your manager who will give you recommendations for your training and the advancement of your career at Inteltrans. Once you have decided what courses to take, proceed to write the email to register for the courses.
INSTRUCTIONS	Now you are ready to reply to that email from Human Resources. Don't forget to specify the course or courses you are taking, the reason and whether you have checked with your manager this training. Below you can find a short list of items you need to address in the email: • Address recipient • Introduce yourself and specify your department • State courses you are planning to take • State whether it's OK for your schedule, from whom you got authorisation, and why you are taking this training • Specify other course(s) you would like to take in the future • Your signature

Table 7.10: Application of the RIF to Activities 1 and 2 in the Final Task in *Education and Training* in ALLES (continues).

Input data



This is your calendar for the current month:

	Monday	Tuesday	Wednesday	Thursday	Friday
9:00-10:00		Emails			Emails
10:00-12:00	Work on current projects			Meeting of Marketing Department	
12:00-13:00	Lunch				
13:00-15:00		Meeting with new partners		Work on reports	
15:00-17:00	Customer service: customer satisfaction reports		Market research for new customers		
17:00-18:00					

Now, read the email from Human Resources and listen to the message from your manager.

From:
To:
Subject:

Table 7.10: Application of the RIF to Activities 1 and 2 in the Final Task in *Education and Training* in ALLES.

As for the relationship between input and response, the activity has a notably broad scope both in the input and in the response. They are both complex in language terms and lengthy. The learner is required to process a considerable amount of information (an audio message, a course list and a calendar page) to extract from them the relevant information. The response to be produced is an email with a considerable amount of information.

As for the directness of the relationship, this is again relatively close in terms of topical knowledge but rather open in terms of the expected language knowledge. The lexical contents to describe the relations between entities that are expected in the response are not guided by the instructions, maybe a bit by the input data, although the instructions include specific information on the email structure.

Table 7.11 includes a detailed analysis in terms of the thematic and the linguistic content of the response to this activity. The first block in the table presents the expected entities and relations:

- Your name and the name of the department in which you work
- The course(s) in which you are interested
- Your availability during the course hours

- The authorisation from your manager for registering
- The profit that you expect to gain from attending the course
- Other courses you would be interested in in the future

Thematic content of the expected response	
Entities	– *I*, as a sender of the request: *my* name – Related to the department: *Marketing Department*, *David Altman* – The names of the courses: *Business Communication* and *E-commerce and E-business.* – Related to time: *schedule*, days of the week, day times, *availability* – Related to manager approval: *manager, approval, permission, ...* – Related to why the course is relevant for the applicant: *improve, improvement, better position, future position, future projects, ...*
Relations	– State the applicant's affiliation to the Marketing Department – State the courses that suit your needs and availability – Argue decision in connection with future plans – State agreement with or approval by your supervisor – Express interest in other courses in the future

Linguistic content of the expected response	
Functional	– Asking others to perform an action – register one for a course – Expressing an interest or an intention – Arguing for decisions – Understanding pieces of advice – Understanding course descriptions – Understanding calendar pages
Syntactic	– Present tense to express current states of affairs – Future tenses (*be* + GERUND, *going to* + INFINITIVE, *will* + INFINITIVE, ...) to express future plans – Causative sentences to express one's own or third-party interests in making decisions – All the syntactic phenomena related with the lexical choices included in the response

Table 7.11: Thematic and linguistic content according to the RIF for Activities 1 and 2 in the Final Task in *Education and Training* (continues).

LEXICAL	*Work (in), Marketing Department, name, register, sign up, attend, course, match, affect, schedule, free*, time expressions, *authorisation, manager, intend, apply, position, career, future, useful, project, take a course*, etc.
PRAGMATICS	Provide information in a coherent and cohesive manner – Use adequate discourse markers and pronouns to glider the text – Email structure: greeting, body, complimentary close, signature
GRAPHOLOGY	– Observe grammar and spelling required in private professional contexts

Table 7.11: Thematic and linguistic content according to the RIF for Activities 1 and 2 in the Final Task in *Education and Training*.

The information to be provided is mostly in the input data: that *you* work in the Marketing Department, that *your* manager's name is David Altman, and that he would be happy that *you* take the course on Business Communication or the one on E-commerce and E-business, and so on.

The second block in Table 7.11 describes the linguistic contents of the response. The task aims at eliciting from the learner linguistic expressions to express a will, to introduce herself/himself and her/his department's name, to argue on the basis of third-party advice's, to express interests, and to justify decisions.

In terms of syntax, learners are expected to use the present tenses to describe states of affairs; to use future tenses to express future intentions and expectations; to use subordination or coordination to express causes; and to use appropriately the functional words – prepositions, conjunctions, etc. – as required by the linguistic items relevant for the response.

As for lexical content, there are several expressions within the semantic field of training courses, career plans, company structure, etc. that learners could use to produce a response. Some of these expressions are found in the input data, but most of them are practised through the unit. In any case, they can hardly be taken as is from the different places where they appear and learners need to resort to their own creative processes to produce the correct linguistic structures.

As for the pragmatic contents, learners are required to provide typical information in emails, some of which is already provided, and ensure that it contains (i) a greeting, (ii) the body of the message containing the expected information, (iii) a complimentary close, and (iv) a signature.

7.2.2.3.2 Assessment

We first present the criteria for correctness, which determine the feedback corresponding to formative assessment, and in the following subsection we present the summative assessment strategy proposed.

A correct response to this task should include:

- An email with addressee, subject – provided as input data – and text
- The text email should contain:
  - Greeting to address colleagues in the human resources dept.
  - Body, with the required information – see above and Table 7.11
  - Complimentary close
  - Signature
- In terms of language knowledge the response has to include:
  - The appropriate expressions to introduce oneself, describe intentions, require actions, report opinions, etc.
  - Use the appropriate syntactic structures in accordance with the lexical choices
  - Structure the text in a coherent and cohesive manner

With all this information a RIF-based sample response can be generated. We present it and comment on it in Section 7.2.2.3.3, after we present the summative assessment criteria for this task.

Summative assessment criteria The criteria for summative assessment are rooted in the pedagogical goals of the unit. The purpose of this assessment in ALLES is that of a low-stakes assessment: to grade and to evaluate learner's progress. To do so, in collaboration with SLA experts in ALLES, we defined indicators based on textual cues that can be used for the assessment of the responses. These indicators correlate with the four aspects we mentioned in Section 6.2.6.2: communicative contents, lexical contents, sentence structure and accuracy, and overall text layout.

These indicators as well as how they combine and link to specific grades and feedback messages are shown in Tables 7.12, 7.13, 7.14, and 7.15. Since these are four very large and wide tables they are presented in landscape orientation after commenting on them, which we do in the following paragraphs. The implementation of the feedback generation strategy of these assessment criteria is described in Chapter 8.

First of all, the tables present two to three columns containing indicators for the linguistic items to be identified for each of the four aspects. In the following two columns, a grade ranging from 4 to 0 is linked to each possible combination of values for each indicator. The different combinations of the values of the indicators are correlated with different grades and different "canned" feedback messages. The feedback that the learner gets results from adding up all the messages obtained for each of the dimensions.

As for the communicative contents, Table 7.12 shows two indicators that take into account the presence of the expected thematic content (TC), as well as the expected linguistic content (LC) at the level of text genre. Indicators are based on the identification of pieces of information (language chunks) that correlate with the expected elements of the response, linguistic or thematic. In this task, elements corresponding to thematic contents are expressions relating to introducing yourself, saying what department you work for, and so on. As for elements corresponding to linguistic content, these are the greeting, the complimentary close, etc.

Using this table, when a response obtains the best possible value for both indicators gets the message "Very good. You use the expected functions adequately" (first row in Table 7.12). In contrast, if a response obtains the best value with respect to the thematic content but the worst one with respect to the linguistic content, then gets a message as "Careful: the exercise has enough contents, but it is not polite" (fourth row).

As for lexical contents, Table 7.13 shows two indicators related to the use of specific vocabulary (SV) and word fluency (the number of words, NW). Assuming a set of reference values for these two indicators, a learner response can be analysed in a way that two values for each of the indicators are associated with the response. The response values are compared to the reference values and the percentage of overlapping between them is reflected in columns SV and NW and linked to specific grades and feedback messages. Again, good indicator values in both SV and NW yield higher grades and more positive messages, but good value indicators in one of the two combined with low values in the other yield less positive messages.

As for sentence structure and accuracy, Table 7.14 shows the three indicators used to assess it: the number of sentences (NS), the number of discourse markers (NDM), and the number of grammar and language usage errors (NGE). This table has absolute values that have been defined by ALLES content designers according to their definition of the task and their experience as FLTL teachers.

The feedback that can be generated thanks to these three indicators are based on formal aspects of written communication. Three sample of messages are: (i) "Great. Your text is correct and adequate. There are no mistakes.", (ii) "Careful, the text is adequate but there are too many errors.", and (iii) "Careful. Your text is adequate but you are not using any connecting words."

As for overall text layout, Table 7.15 shows the indicators used to assess it: the number of paragraphs (NP) and the number of spelling errors (NSE) in the whole response. As it happens in Table 7.14, Table 7.15 presents absolute figures that have been defined by ALLES contents designers on the basis of their experience. Value combinations yield different sorts of feedback messages.

Communicative contents

TC	LC	Grade	Message
6	4	4	Very good. You use the expected functions adequately.
6	3	3	Very good. You use almost all of the expected functions adequately.
6	2	3	Good. Although you adequately use the expected functions, review the courtesy.
6	1	1	Careful: the exercise has enough contents, but it is not polite.
5	4	3	Very good. You use almost all of the expected functions adequately.
5	$3 \leq x \leq 2$	2	Good. Although you use almost all of the functions, review the courtesy.
5	1	1	Careful: there is some information missing in the exercise.
4	4	2	Careful: there is some information missing in the exercise.
4	$x \leq 3$	1	Are you sure you have understood the purpose of this exercise?
3	$x \leq 4$	0	Are you sure you have understood the purpose of this exercise?

Table 7.12: Indicators for the assessment of communicative contents: thematic content (TC) and linguistic content (LC) at the level of text genre.

Lexical contents

SV	NW	G	Message
$80\% \leq 100\%$	$90\% \leq 100\%$	4	Excellent. Your text reads well and is precise. You are using the (...)
$80\% \leq 100\%$	$50\% \leq 89\%$	3	Good. Your text is pertinent but you should be more fluent.
$80\% \leq 100\%$	$0\% \leq 49\%$	2	Careful: You are using adequate vocabulary but the text does not read well.
$50\% \leq 79\%$	$90\% \leq 100\%$	3	Excellent. Your text reads well, but you should use specific vocabulary.
$50\% \leq 79\%$	$50\% \leq 89\%$	2	Try to be more fluent and use specific vocabulary.
$50\% \leq 79\%$	$0\% \leq 49\%$	1	Careful: your text does not read well and you should use more (...)
$30\% \leq 49\%$	$90\% \leq 100\%$	2	Good. Your text reads well, but you should use specific vocabulary.
$30\% \leq 49\%$	$50\% \leq 89\%$	1	Careful; Try to be more fluent. Check the vocabulary you are using.
$30\% \leq 49\%$	$0\% \leq 49\%$	0	Careful. Try to write a text that reads well. Check the vocabulary.
$0\% \leq 29\%$	$90\% \leq 100\%$	1	Good. Your text reads well, but you should use specific vocabulary.
$0\% \leq 29\%$	$50\% \leq 89\%$	1	Careful; Try to be more fluent. Check the vocabulary you are using!
$0\% \leq 29\%$	$0\% \leq 49\%$	0	Careful. Your vocabulary is inappropriate and the text does not read well.

Table 7.13: Indicators for the assessment of lexical contents: use of specific vocabulary (SV) and the number of words (NW) as a fluency measure.

Sentence structure and accuracy

NS	NDM	NGE	G	Message
$10 \leq x \leq 9$	$10 \leq x \leq 9$	$0 \leq x \leq 1$	4	Great. Your text is correct and adequate. There are no mistakes.
8	$10 \leq x \leq 9$	$0 \leq x \leq 1$	4	Great. Your text is adequate, but there are some minor errors.
$7 \leq x \leq 6$	$10 \leq x \leq 9$	$0 \leq x \leq 1$	3	Good. But some information is missing.
$5 \leq x \leq 0$	$10 \leq x \leq 9$	$0 \leq x \leq 1$	2	Careful. Your text is too short or has too long sentences, though it is adequate.
$10 \leq x \leq 9$	$10 \leq x \leq 9$	2	3	Great. Your text is adequate, but there are some minor errors.
8	$10 \leq x \leq 9$	2	3	Good. Your text is adequate, but there are some minor errors.
$7 \leq x \leq 6$	$10 \leq x \leq 9$	2	2	Good. But some information is missing, and there are some minor errors.
$5 \leq x \leq 0$	$10 \leq x \leq 9$	2	2	Careful. Your text is too short or has too long sentences, and it has some grammatical mistakes.
(...)				
$10 \leq x \leq 9$	$10 \leq x \leq 9$	$x \geq 5$	3	Careful, the text is adequate but there are too many errors.
8	$10 \leq x \leq 9$	$x \geq 5$	2	Careful, the text is adequate but there are too many errors.
$7 \leq x \leq 6$	$10 \leq x \leq 9$	$x \geq 5$	2	Careful, some information is missing, and there are too many errors.
$5 \leq x \leq 0$	$10 \leq x \leq 9$	$x \geq 5$	1	Careful, the text is not adequate and there are too many errors.
$10 \leq x \leq 9$	$8 \leq x \leq 6$	$0 \leq x \leq 1$	3	Good. Your text is adequate, but there are some minor errors.
8	$8 \leq x \leq 6$	$0 \leq x \leq 1$	3	Good. But some information is missing. Check it out.
$7 \leq x \leq 6$	$8 \leq x \leq 6$	$0 \leq x \leq 1$	2	Careful; some information is missing and there are some minor errors.
$5 \leq x \leq 0$	$8 \leq x \leq 6$	$0 \leq x \leq 1$	1	Careful, some information is missing and the text is not adequate.
(...)				

Table 7.14: Indicators for the assessment of sentence structure and accuracy: number of sentences (NS), number of discourse markers (NDM), and number of grammar and usage errors (NGE) in the response (continues).

8	$9 \leq x \leq 10$	$x \geq 5$	2	Careful: there are global and local syntactic problems in your text. Check it out, please.
8	$6 \leq x \leq 8$	$x \geq 5$	1	Careful: there are global and local syntactic problems in your text. Check it out, please.
8	$6 \leq x \leq 5$	$x \geq 5$	1	Careful: you are not using any connecting words. There are too many errors as well.
$8 \leq x \leq 10$	$x \leq 5$	$0 \leq x \leq 2$	1	Careful. Your text is adequate but you are not using any connecting words.
$6 \leq x \leq 7$	$\forall$ x	$x \geq 5$	0	Careful: some information is missing. There are too many mistakes as well.
$5 \leq x \leq 0$	$6 \leq x \leq 10$	$x \geq 3$	1	Careful. Some information is missing and the text has some grammatical mistakes.
$0 \leq x \leq 7$	$x \leq 5$	$x \geq 3$	1	Careful. Some information is missing and you are not using any connecting words.
(...)				

Table 7.14: Indicators for the assessment of sentence structure and accuracy: number of sentences (NS), number of discourse markers (NDM), and number of grammar and usage errors (NGE) in the response.

Overall text layout

NP[2]	NSE	G	Message
9	$0 \leq x \leq 1$	4	Excellent. Your text has an adequate structure and no spelling mistakes.
9	$2 \leq x \leq 3$	2	Good. Your text has an adequate structure, but some spelling mistakes.
9	$x \geq 4$	1	Careful. Your text has an adequate structure but there are many spelling mistakes. Check them, please.
$6 \leq x \leq 8$	$0 \leq x \leq 1$	3	Your text has a somewhat adequate structure and no spelling mistakes.
$6 \leq x \leq 8$	$2 \leq x \leq 3$	1	Your text has a somewhat adequate structure and some spelling mistakes.
$6 \leq x \leq 8$	$x \geq 4$	0	Your text has a somewhat adequate structure but you should check spelling.
$x \leq 5$	$0 \leq x \leq 1$	3	Your text does not have any structure. There are no spelling mistakes.
$x \leq 5$	$2 \leq x \leq 3$	1	Your text does not have any structure. There are some spelling mistakes.
$x \leq 5$	$x \geq 4$	0	Careful: Your text does not have any structure and has many spelling errors.

Table 7.15: Table co-relating the number of paragraphs (NP) and the number of spelling errors (NSE) in the response.

[2] Though NP stands for number of paragraphs, for this particular activity, content designers propose to take into account the number of sentences arguing that the text is too short for paragraph counting.

7.2.2.3.3 A RIF-based list of predictable responses

With the RIF analysis, a sample response can be built. Again, however, instead of building our own list of predictable responses, we use a sample response provided by one of the FLTL practitioners involved in material creation.

(9) Dear Madam, Dear Sir,
My name is NAME and I work in the Marketing Department of Inteltrans. I have signed up to do the Business Communication course. It does not affect my schedule as I am free on Monday, Wednesday and Thursday mornings. I am taking this course with the permission of my manager because I intend to apply for a more senior position within the company later on this year and I believe that this course will help me for my project. In the future, I'd like to take the E-commerce and E-business course.
Best Regards,
SIGNATURE

From the sample answer in (9), some more interesting characteristics of the response to this task can be identified.

At the level of thematic contents, all the information provided matches with the instructions and the restrictions set by the input data except when it says "because I intend to apply for a more senior position within the company later on this year and I believe that this course will help me for my project". However, the manager's voice mail, part of the input data, recommends the employee (the learner) to take two courses because "they could be useful for the marketing projects we will have to develop by the end of the year" (see voice mail transcript in Section C.3.1 in Appendix C).

The content designer's choice might be accepted as correct since the task's instructions are not too restrictive in this aspect. The relevant passage in the instructions reads "why you are taking this training". Despite this, one could justify taking the training on the basis of the manager's words, the sample response extends that information with a reasonable proposition "because I intend to apply for a more senior position within the company later on this year". This implies however that it will be harder for the NLP tools to provide a reliable analysis that can ensure or discard the presence and appropriateness of this specific information in the learner response – further details in Sections 8.4.3 and 8.4.4.

Other aspects to be inferred from this sample answer are: (i) that the email can be started with *Dear Madam, Dear Sir* – and then one should think about other possible openings–; (ii) that the learner can decide to register for just one course – instead of registering for the two courses that her/his manager recommends and that fit with her/his schedule; and (iii) that lexical choices other than the ones mentioned in the instructions and input data can be found, such as *sign up* as a synonym for *register*, *permission* as synonym for *authorisation* – which is used in the input data –, or *believe X will help me* as a structure to argue for the appropriateness of a decision. As it happened with the task exemplifying Task type II, by analysing the sample response an expansion of the RIF definition is possible, one that will offer new possible alternative correct responses.

7.2.2.4 Task type IV

The task that exemplifies Task type IV corresponds to Activity 5 of Subtask 2 in *Atención al cliente*, the Spanish version of Customer Service and International Communication in ALLES. The task's title is *Expresa tu satisfacción o insatisfacción con el producto* Smint.[3]

The TAF-characterisation of the task is:

Description Writing task that requires the learner to put herself in the shoes of a Smint consumer, a candy; the consumer sends a letter to the manufacturer to express her opinion on the candy and to ask for further information.

Focus Meaning and form.

Outcome A consumer letter.

Processes Expressing positive and negative aspects of consumer products, particularly candies; asking for information related to consumer products.

Input The learner is provided with a prompt describing the fictive setting and is given some hints on how to ask for information.[4]

Response type Extended written production: a letter.

Teaching goal Structured communication.

Assessment Formative.

This writing task focuses on meaning with a communicative outcome. Again, a role-play activity in which the learner is put in the place of a consumer that writes a letter expressing an opinion about a product. The processes that underlie this learning activity include processes such as giving an opinion about a product or asking for information.

The task requires an extended production response, a letter. The type of learning task, the teaching goal, could be classified as structured communication and the assessment it requires, not being a final task, is formative assessment.

7.2.2.4.1 Characterisation of the response in pedagogical and linguistic terms

Table 7.16 characterises the response to this task in terms of instructions and input. The task's prompt requires the learner to put himself or herself on the role of someone who is invited to send her/his opinion on *Smint*, produced by the Spanish

[3]The title in English would be: *Express your satisfaction or dissatisfaction with the product* Smint.

AC, Subtarea 2, Actividad 5

Prompt	Imagina ahora que tú también has participado en la encuesta de satisfacción para conocer la aceptación del nuevo producto de Chupa Chups SA, SMINT, y quieres escribir una carta a la empresa para expresarles directamente tu opinión tras probar los caramelos. Además, quieres aprovechar la circunstancia para pedir más información sobre el producto.
Instruction	En el cuadro de abajo puedes consultar algunas estructuras de cómo pedir información.
Input data	Cómo “PEDIR INFORMACIÓN” Aquí tienes algunas estructuras que puedes utilizar para solicitar información: – ¿Podría/puede decirme si ...? – ¿Quería saber si ...? – Me gustaría saber si ... Ejemplos: – Por favor, ¿podría decirme si el tren que va a Zaragoza tiene parada en Lleida? – ¿Quería saber si los estudiantes tenemos descuento en los museos? – Me gustaría saber si la próxima semana podemos visitar la nueva fábrica para ver todos los adelantos técnicos que se han incorporado.

Table 7.16: Application of the RIF to Activity 5 in Subtask 2 in *Atención al cliente.*

company *Chupa Chups, SA*. The learner is encouraged to send a letter and to seize the opportunity to ask for further information.

The prompt assumes that the learner has gone through the previous tasks in Subtask 2 in *Atención al cliente*, where s/he is introduced to the topic by reading a corporate report summarising customer satisfaction for *Smint*, as well as some consumer opinions. In the previous tasks the learner was introduced to some linguistic resources useful to express satisfaction and dissatisfaction. The prompt and the instructions do not pose any further restriction on the response.

As for the input data, the learner is provided with a space to respond and a list of formulas frequently used in Spanish to ask for information. For each of these formulas an example is provided. Examples are not related to the task's topic.

As for the relationship between input and response, this task presents a considerably broad scope. The length of the instructions and the input is relatively short, and the length of the response is open. They need to produce a letter expressing their opinion, as much as they need to say, as little as they need to say.

As for the directness of the relationship between input and response, this is relatively indirect. To complete it, learners can rely on non-linguistic and linguistic resources of their own choice. It is not a free composition. The topic is a specific candy and it has to be about expressing likes and dislikes, and asking something about it. But the distance between input and response is big. The actual contents of the response are not much restricted by the input. Learners are expected to express opinions and ask for information, and are given some hints on how to do it, but not required to use specific expressions.

The thematic and the linguistic contents for this task are characterised in Table 7.17. The first block in the table presents the entities expected in the response: there has to be a reference to the product *Smint*. There might be references to entities such as *sweets*, *candies*, *lollipops*, or *product*, and even to entities such as *company*, *Chupa Chups, SA*. But the reason that made consumers happy or unhappy with *Smint* is uncertain, or open. It could be the price, the size, the flavours, the colour, the packaging, etc.

As for the relations, instructions impel the learner to state that s/he bought or tried Smint. Anything else related to satisfaction with *Smint* depends on the learner's background or imagination. The instructions require learners to ask about the product. However, might it be about whose idea was it to pack it in such a box? Or about what flavour is to appear next in the market? Or what else? Answers are open in terms of thematic content, and difficult to predict in lexical terms.

As for the analysis in terms of linguistic contents, instructions are explicit about part of the expected functional contents: ask for information. The linguistic input data below the space reserved for the answer give specific formulas for this.

As for the expression of opinion, only the previous tasks in the Subtask might help. Learners have to express their opinion about a candy, a type of food. This will include describing it, being able to say positive and negative things about it, and so on. The exponents of function provided to learners in previous activities to express likes and dislikes are in the functional contents area in Table 7.17.

In terms of syntactic contents, the response might include past tense to express

Thematic content of the expected response	
Entities	– Smint – Chupa Chups – Sweets, products – Consumer (the learner role)
Relations	– You buy or bought Smint and tried it. – You like and disliked certain things about Smint – You want to know more about Smint

Linguistic content of the expected response	
Functional	– Express what you think about a product: *Me gusta mucho/muchísimo SN/que...* *Me encanta/encantan...* *¡SN está Adj. Superlativo!* *¡Qué malo/bueno (que está) (SN)!* *No me gusta (nada) SN/que...* *Odio SN/que...* – Ask for information: ¿Podría/puede decirme si ...? ¿Quería saber si ...? Me gustaría saber si ...
Syntactic	– Past tense to express what you did – Present tense to explain what you think – Conditional tense (to express preferences or to ask for information) – Courtesy forms in pronouns and verbs (3rd person singular) – Use of relevant prepositions and conjunctions (3rd person singular)
Lexical	– *Smint, Chupa Chups, caramelo, caramelo de palo, gustar, encantar, odiar, sabor, rico, malo, bueno, caro, práctico, sano, etc.*
Pragmatics	– Use the appropriate register. – Use an appropriate letter structure
Graphology	– Use the appropriate spelling.

Table 7.17: Thematic and linguistic content according to the RIF for Activity 5 in Subtask 2 in *Atención al cliente*.

experiences, present tense to express opinion, conditional tense to express preferences or to be polite, and the use of courtesy forms because of the formality of the setting. Also the use of prepositions and conjunctions with topic relevant words will be assessed.

As for lexical content, not all of the expected elements are easy to predict. Instructions and the task topic ensure the appearance of *Smint, candy, Chupa Chups, (no) me gusta, me encanta, odio, ...*. Moreover, talking about sweets might facilitate the use of words such as *rico* [yummy], *sabroso* [tasteful], pegajoso [sticky], sano [healthy], azúcar [sugar], *diente* or *muela* [tooth, back tooth], and similar. However, instructions and input data are open enough in this respect to make it diffcult to predict.

Finally, in terms of pragmatics and graphology, the letter that learners are expected to write has to comply with the norms of the communicative setting and the text genre. If it is a letter, it should include the addressee and sender contact details, a place and a date, usually at the end. It will also require a greeting, the contents of the body (opinion and asking for information), a complimentary close and a signature. Given the relative formality of the communicative setting, the text should observe the courtesy, spelling, grammar, and norms socially required.

7.2.2.4.2 Assessment

Since this task requires formative assessment, the criteria for correctness need to be established in terms of correct/incorrect, so that feedback is provided appropriately. However, correctness for this task is harder to establish than the previous ones in terms of the thematic content: There is no one particular aspect of the product to be praised, nor one to be criticised. A correct response should consist of:

- A letter expressing an opinion about *Smint* with addressee and sender contact details, greeting, body, complimentary close, and signature.
- The body of the letter with:
  - Statement of have tasted *Smint*
  - Opinions about the product *Smint*
  - Questions about the product *Smint*
- In terms of language knowledge the response has to include:[5]
  - Expressions to communicate personal experience with a product
  - Expressions to show satisfaction or dissatisfaction with a product
  - Expressions to ask information about products
  - Structure the text in a coherent and cohesive manner
  - Use expected text type structure
  - Use expected register, as well as appropriate spelling and grammar norm

[5]All these aspects have been described in detail above and are reflected in the corresponding sections in Table 7.17.

7.2.2.4.3 A RIF-based list of predicted responses

Despite producing the criteria for correctness, we do not provide a sample response for this task. None of the content designers provided one during the ALLES project, and its thematic contents are too open. Later on, in Chapter 8, we propose a particular approach to assess this kind of task, and in Chapter 9 we analyse some real learner responses to this task, which we can compare with our initial analysis.

7.3 Chapter summary

In this chapter, we presented two frameworks that we consider an essential part of what should become a methodology for the design and development of ICALL tasks. In our view, these two frameworks establish a connection between the needs and requirements of FLTL, and the linguistic and assessment specifications for NLP. This connection is made in a detailed characterisation of learner responses and assessment criteria in the design phase, as a primary input to NLP.

The TAF and the RIF help describe the pedagogical properties of FL learning activities using concepts and terminology generally accepted in FLTL and SLA. They also help us specify in detail the linguistic contents and the assessment criteria of these tasks using concepts and terminology from the field of linguistics, a field that serves as a common language for FLTL and NLP, both of which are concerned with language but with different perspectives. FLTL looks at language as a communication system that has to be learned and used in a competent manner, while NLP looks at it as a complex system to be computationally formalised.

To exemplify the use of the TAF and the RIF, we applied them to four FL learning activities that exemplify four learning task types, three of which are good representatives of the viable processing ground. Figure 7.1 shows how these four tasks could be approximately situated on the viable processing ground (extending the figure in Bailey and Meurers (2008: p. 108)).

Task type I is most to the left, because its responses are limited production responses, and the interrelationship between input and response is narrow and direct. Task types II and III are slightly further to the right because their responses are extended production responses, the interrelationship between input and response is relatively broad, and for both of them the input-response interrelationship is notably direct both in terms of thematic content, and a bit less in terms of linguistic content. Task type III is different from Task type II in that it requires summative assessment. Finally, Activity type IV is the one that is most to the right, because it presents extended production responses, and a broad and indirect input-response interrelationship, particularly in terms of content.

Once FL learning activities are characterised with the TAF and the RIF, the analyses inform the feedback generation strategy about the pedagogical needs, the nature of the expected responses and the assessment criteria. With this information software specifications can be developed and the implementation of the NLP-based analyses modules started. This is what we focus on in the following chapter.

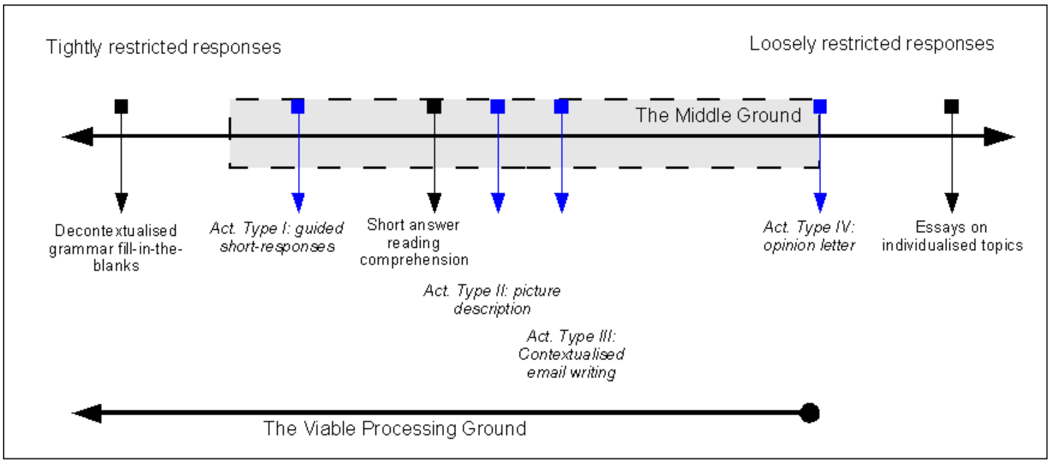


Figure 7.1: Four task types in the viable processing ground.

Chapter 8

NLP functionalities to respond to FLTL demands

This chapter presents a proposal for the specification and implementation of an automatic assessment (AA) module on the basis of automatic linguistic analysis. Because the automatic assessment module we are aiming for is a pedagogically informed one, we propose to base the specifications for the AA module on the RIF, presented in the previous chapter. The RIF provides the pedagogical and linguistic information, sets of expected responses and criteria for correctness, so that we have access to an explicit representation of the information upon which the language processing and feedback generation modules can be based.

We first introduce the Automatic Assessment Specification Framework (AASF), designed for the specification of the technical requirements for both the language analysis module and the feedback generation module. The AASF is step-wise conversion of pedagogical and linguistic features into specific analysis needs at the level of linguistic knowledge and the level of topical knowledge, using Bachman and Palmer (1996)'s terms, or at the level of meaning and form, using Bailey and Meurers (2008)'s terms. We first introduce the AASF formally and then exemplify its application on a selection of ICALL tasks, including both tasks requiring formative feedback and tasks requiring summative feedback.

Second, we present the actual implementation of an automatic assessment module on the basis of the specifications derived from the AASF. Our description includes a general approach to feedback generation on the basis of the shallow NLP processing tools presented in Chapter 6. After this, we describe the implementation of NLP processing strategies to handle task-specific linguistic analysis beyond the morphosyntatic level under the assumption of the task characteristics as a domain. Moreover, we describe how this general approach to feedback generation can be instantiated for specific ICALL tasks, where we distinguish between the approach followed for the generation of formative feedback and the approach followed for the generation of summative feedback.

8.1 From pedagogical requirements to specifications for the Automatic Assessment

This section presents the **Automatic Assessment Specification Framework** (AASF) as a means to establish the requirements of the linguistic analysis strategy and the feedback generation strategy on the basis of the RIF.

8.1.1 AASF: Automatic Assessment Specification Framework

The AASF consists of two main parts: the **Specifications for Automatic Linguistic Analysis** (SALA), which will provide NLP-oriented specifications for the analysis of learner responses to a particular activity, and the **Specifications for the Feedback Generation Logic** (SFGL), which will provide a feedback generation logic to make hypotheses on the correctness of learner responses that link the linguistic analysis to "canned" feedback messages.

8.1.1.1 SALA: Specifications for Automatic Linguistic Analysis

The linguistic analysis pursues both the analysis of meaning and the (relevant) analysis of form. The analysis of meaning will be related to the thematic content specifications of the RIF, where the topical knowledge of the activity is characterised. As for the analysis of form, it will be related to the linguistic content specifications of the RIF, where the linguistic knowledge of the activity is characterised.

For each item in the thematic and linguistic content parts of the RIF, the SALA provides the following information:

1. **Reference**: a description of the individual or the relation that a particular piece of language will be referring to

2. **Linguistic cues**: specific linguistic units and structures that can be associated to the *reference*

3. **Code**: a codification for the analysed phenomenon to be linked with particular feedback generation logics

4. **NLP module**: the NLP module or functionality required to trigger the expected piece of language

Figure 8.1 reflects the procedure through which the elements of the thematic and linguistic content parts in a RIF analysis of an ICALL activity can be converted into NLP specifications. From step one to step two, for each of the items in the RIF specifications a set of cues has to be defined in terms of linguistic or textual information. The third step consists in establishing a code for the phenomenon on which the logic for feedback generation will be based. The last step is to identify the NLP module that can provide the function, the automated procedure, to analyse (detect, annotate) the expected linguistic cues.

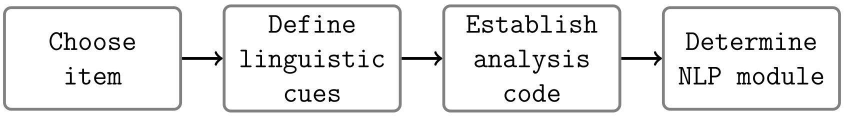


Figure 8.1: NLP specification procedure.

8.1.1.2 SFGL: Specifications for the Feedback Generation Logic

The feedback generation logic is the brain – the "reasoning" – of the Feedback Generation module. It pursues the assessment of the learner response in terms of criteria concerned with a reduced number of words in the response, local level, or with the response as a whole, global level. Independent of the locality or globality of the criterion, assessment aims at the meaning and the form aspects of the response, which are related to the thematic and linguistic contents in the RIF.

For each of the items in the RIF-based criteria for correctness, the SFGL provides the following information:

1. **Criterion**: the particular criterion to be checked for in the learner response
2. **Priority**: the criterion's priority can be high, medium or low
3. **Match**: the type of match is defined as full, partial or zero match
4. **Message**: the feedback message associated with each type of match
5. **Type**: messages can be mutually exclusive, marked as *main*, or can be added to messages of type main, marked as *addable*.

Figure 8.2 reflects the procedure for the specification of the feedback generation logic. The first step is to select the criterion, or the conditions (codes or code sequences), for which specific feedback messages will be generated. Afterwards the priority of the criteria can be established. A higher priority will result in a greater prominence in the feedback presentation. A feedback message for the three different matching options (full, partial, and zero) can be written. The last step is to determine for each the feedback messages whether they have the type main or the type addable.

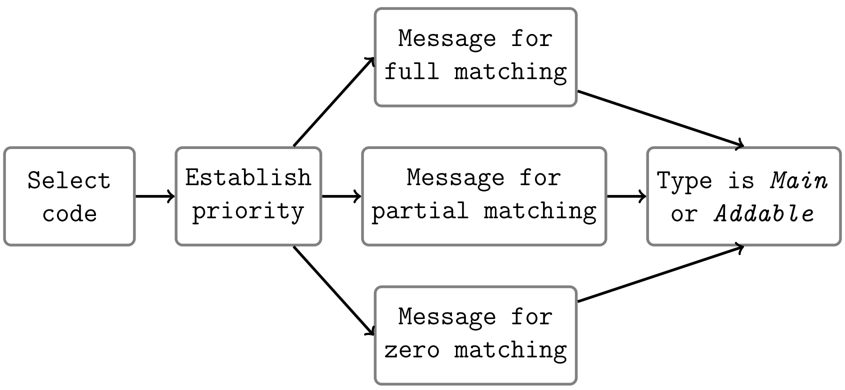


Figure 8.2: Specification procedure for the feedback generation logic.

8.2 Applying the AASF to ICALL tasks

In this section we exemplify the application of the AASF to tasks requiring formative feedback, as well as tasks requiring summative feedback. As we will see, both the SALA and the SFGL are applied to the topical and the linguistic knowledge of the response. For simplicity, their application to each type of knowledge will be shown in separate tables.

8.2.1 Applying the AASF for formative assessmemt

This section presents the application of the AASF for the specification of the automatic assessment module of the ICALL task Activity 4 in Subtask 3 in the learning unit *Customer Satisfaction and International Communication*.

8.2.1.1 The SALA applied to formative feedback

Tables 8.1 and 8.2 show the results of applying the SALA to Item 1 in the above mentioned activity. Tables 8.1 is the application of the SALA in terms of thematic contents in the RIF, and Table 8.2 is its application in terms of linguistic contents (see details in Section 7.2.2.1).

Table 8.1 has four columns, each of them corresponding to one of the items of the SALA as we just described. Moreover, Table 8.1 is divided in two larger areas labelled with the terms ENTITIES and RELATIONS, the two divisions of the thematic contents in the RIF.

As for entities, in the first column there are the references to the interviewer, the interviewee, and the service in question, *the Stanley Broadband service*. Each of them has an associated set of linguistic cues. For this particular activity, the interviewer has no explicit linguistic cues (or then the question as whole), the interviewee has a single word ('you'), and the service in question is stated with complex linguistic expression ('Stanley Broadband').

In the third column, an analysis code is established for the identification of the corresponding text sequence in the learner response when the reference is properly analysed. In the fourth, a corresponding NLP technique is identified as most appropriate. The detection of the entity interviewee requires being able to detect the presence of a second-person singular personal pronoun, namely *you*. For this a lemmatisation and a POS tagging process is required. The detection of the entity *Stanley Broadband* requires a Named Entity Recognition module.

As for relations, in the lower block of Table 8.1, the only reference is the one concerned with asking the interviewee about his/her level of satisfaction with the service. As shown in the second column of the last row, this relation can be realised linguistically with certain lexical combinations: *how satisfied are YOU with X* or *how did YOU enjoy using X*. Since the textual cues are extracted from the RIF analysis, any expression of level of satisfaction that does no start with *how* is excluded. If there was the need to handle responses that do not (strictly) follow the instructions, synonymous expressions using differing structures could be specified – such as *what is your level of satisfaction with X*.

Reference	Textual cue	Analysis code	NLP module
Entities			
The interviewer	- the question as a whole rather than a linguistic cue referring to him/her	–	–
The interviewee	- *you*	Lemma:YOU	Tagger with lemmatisation
The Stanley Brodband service	- *Stanley Broadband* - *the Stanley Broadband service*	NE:SBService NE:SBService	Named Entity Recogniser Named Entity Recogniser
Relations			
Level of satisfaction of interviewee	- *how satisfied are YOU with X* - *how did YOU enjoy using X*	Rel:LevelOfCustomerSatisfWith Rel:LevelOfCustomerSatisfWith	IE module (relations) IE module (relations)

Table 8.1: SALA applied to the thematic content part of the RIF for Item 1 of Activity 4 in Subtask 3 of the Learning Unit *Customer Service and International Communication* in ALLES.

Table 8.2 has the same four columns and is divided in five larger areas labelled FUNCTIONAL CONTENTS, SYNTATIC CONTENTS, LEXICAL CONTENTS, PRAGMATIC CONTENTS, and GRAPHOLOGY – which correspond to the parts of the linguistic content analysis in the RIF. The process is exactly the same: For each of the selected criterion a set of textual cues is defined, a code is assigned to them and an NLP module or functionality is identified.

As shown in Table 8.2, the functional contents partially overlap with the specifications in the thematic content. The difference lies in that here the goal is to identify that the learner produces the language required for a communicative function to take place, while, in the thematic content, the goal is to identify that the learner communicated *the information* he or she was *expected* to communicate in *that* communicative setting.

As for syntactic contents, the criteria for correctness in the RIF require for this activity to detect whether an interrogative sentence starts with *how* and whether it ends with an interrogation mark. As the table shows, this kind of form analysis does not necessarily require a parser, a tokenisation and lemmatisation functionality suffice. In contrast, the analysis of subject-predicate inversion does require some sort of syntactic analysis, therefore a syntactic parser is required.

As for the lexical contents and the register (in the pragmatic contents part), their analysis is dependent on communicative setting – at least for the four languages worked with in this research. The politeness or the formality associated with a word depends highly on the word itself. It can be in its lemma as in these two English expressions for greeting: *hi-you* vs. *good morning*; or in its morphological information as in the Spanish verb forms *canta* [sing IMPERATIVE-2ND-PERSON SINGULAR-COLLOQUIAL] or *cante* [sing IMPERATIVE-3RD-PERSON SINGULAR-POLITE]. Therefore, we require the use of the lexicon and the syntactic parser.

Finally, the section devoted to graphology might require language checking, so a spell and grammar checker are required.

Reference	Textual cue	Analysis code	NLP module
Functional contents			
Asking somebody about his or her satisfaction with something	- how happy/satisfied are you with	Func:AskAboutSatisfaction-With	IE module (exponents)
	- how much did you enjoy using X	Func:AskAboutSatisfaction-With	IE module (exponents)
Syntactic contents			
Interrogative sentences	- Starts with *how*	Synt:HowFirst	Lemmatisation and sentence position
	- Ends with "?"	Synt:IntMarkLast	Tokenisation
	- Subject-predicate inversion	Synt:SubjPredInversion	Syntactic parser
Lexical contents			
Use of relevant content/functional words	- *how*, *be satisfied with*, *Stanley Broadband*, *service*, ...	Lemma	Lexicon
Pragmatic contents			
Appropriate register	- Appropriate formal expressions	RegisterOk	Lexicon and syntactic parser
Graphology			
Appropriate spelling	- Word spelling and punctuation	SpellingOk	Spell/Grammar checker

Table 8.2: SALA applied to the linguistic content part of the RIF for Item 1 of Activity 4 in Subtask 3 of the Learning Unit *Customer Service and International Communication* in ALLES.

8.2.1.2 The SFGL applied to an activity with formative feedback

With the requirements for the automatic linguistic analysis, the feedback generation logic can be specified. The analysis codes are the link between the SALA specifications and the SFGL specifications. The application of the SFGL to linguistic and content knowledge is shown in separate tables.

Tables 8.3 and 8.4 present the result of applying the SFGL to Item 1 of the task we are exemplifying (see Section 7.2.2.1). In both tables, each of the five columns corresponds with the five items of the SFGL. The first column shows the criterion, that is, the analysis condition that triggers a particular feedback; the second, the priority of the condition; the third, the type of matching (full, zero or partial); the fourth, the message itself; and the fifth, whether it is a message that can combined with messages related to other criteria or not.

Table 8.3 presents two blocks, one for the assessment logic related to ENTITIES and another one for the logic related to RELATIONS. All the analysis conditions in both blocks are marked as *High* in terms of priority, since they are related to aspects crucial for the fulfilment of the criteria for correctness of the response.

In the ENTITIES block, assessment is only associated with particular messages in case the matching is zero, that is, in case, the response has no reference to any of such two entities. Both messages are assigned the type *Addable*, so they might co-appear with other messages. The reason for this responds to the will to give more prominence to assessment relating to the response as a whole than to smaller "bits" of the response (see Table 8.4 and related comments below).

As for the RELATIONS block, the first row presents the assessment specification that will check the complete expected contents of the response: *Rel:LevelOfCustomerSatisWith + NE:SBService*. That is: the customer is being asked about his/her level of satisfaction with the Stanley Broadband service, two of the analysis codes defined in the NLP specifications checked for in a row. For each of the matching possibilities a different message is foreseen, and all are of the type *Main*.

A for the second row, it contains the logic to check for the situation in which the customer is asked about his/her level of satisfaction with something that was not detected as the entity 'Stanley Broadband'. This analysis condition has high priority, and its messages are of type Main. Messages for almost all matching conditions exist.

The very last row indicates that the feedback generation specifications can be as long as particular relevant cases can be thought of. The main issue in this respect is to balance the effort-benefit costs of defining particular analysis/feedback generation strategies to provide more fine-grained feedback.

Analysis condition	Prio.	Match	Message	Type
Entities				
Lemma:YOU	High	Full	–	–
		Zero	Are you addressing your customer? I cannot find a reference to him/her.	Add
		Partial	–	–
NE:SBService	High	Full	–	–
		Zero	Are you mentioning the name of the service? I cannot understand it.	Add
		Partial	–	–
Relations				
Synt:HowFirst + Rel:LevelOfCustomer-SatisfWith + NE:SBService + Synt:IntMarkLast	High	Full	Good! Your answer is correct!	Main
		Zero	Your answer is not appropriate. Try again!	Main
		Partial	Your answer is correct in terms of contents.	Main
Rel:LevelOfCustomerSatisfWith + Unknown	High	Full	Your answer is partially correct. A reference to X is missing.	Main
		Zero	NA	NA
		Partial	Your answer is partially correct. It has some language error and a reference to X is missing. Try again!	Main
(...)				

Table 8.3: SFGL applied to the thematic content part of the RIF for Item 1 of Activity 4 in Subtask 3 of the Learning Unit *Customer Service and International Communication* in ALLES.

Table 8.4 presents the feedback generation logic for the assessment of linguistic content, which is considered relevant for the task response according to the RIF. All analysis conditions are qualified as high priority. The table is divided in the five relevant assessment areas found for the linguistic content analysis in the RIF of this activity.

As for FUNCTIONAL CONTENT, no particular condition is specified. This reflects the fact that for this response there is no difference between having expressed correctly the expected thematic contents and using the expected exponents for the function. If for some reason, there was any exponent for the function "ask about satisfaction with" for which specific feedback messages were to be generated, then it would be added here.

As for SYNTACTIC CONTENT in Table 8.4, it presents messages for three assessment criteria defined in the RIF. Note they are all of the type *Addable*, and that they only present a message for the zero matching condition. Positive reinforcement messages could be added to full matching condition, and other kind of messages could be added in case of partial matching – and what exactly is meant by partial matching in each of the three cases would have to be defined.

As for the LEXICAL CONTENT part, there are messages for the cases that none or only a part of the relevant words expected in the response are found in the learner's response. Here what the relevant words for the response are has to be defined, of course. In this case, we chose the words *how*, any word that could be used to express satisfaction (*to be satisfied/happy with*, *to enjoy*, etc.) and the name of *the Stanley Broadband service* in some form.

As for the parts relating to PRAGMATIC CONTENT and GRAPHOLOGY, they present messages for all matching conditions. Note that they are of the type *Addable*, since these are messages that do not inform of the correctness of the response in terms of contents and would make little sense, pedagogically speaking, on their own.

Analysis condition	Prio.	Match	Message	Type
Functional content				
Func:AskAboutSatisfactionWith	High	Full	–	–
		Zero	–	–
		Partial	–	–
Syntactic content				
Synt:HowFirst	High	Full	–	–
		Zero	Your answer should start with *how*.	Add.
		Partial	–	–
Synt:IntMarkLast	High	Full	–	–
		Zero	Your answer should end with a question mark.	Add.
		Partial	–	–
Synt:SubjPredInversion	High	Full	–	–
		Zero	Your answer does not seem to follow the word order of questions in English.	Add.
		Partial	–	–
Lexical content				
Lemmas of concept words expected	High	–	–	
	High	Zero	The answer is not appropriate.	Main
	High	Partial	The answer contains some of the relevant words, but is not quite what I expected.	Main

Table 8.4: SGFL applied to the linguistic content part of the RIF for Item 1 of Activity 4 in Subtask 3 of the Learning Unit *Customer Service and International Communication* in ALLES (continues).

PRAGMATIC CONTENT				
RegisterOk	High	Full	You are using the appropriate register.	Add.
		Zero	Be careful, your text might result impolite.	Add.
		Partial	Be careful, part of your text might result impolite.	Add.
GRAPHOLOGY				
SpellingOk	High	Full	Your text has no language errors.	Add.
		Zero	Be careful, your text has too many incorrections.	Add.
		Partial	Be careful, your text has some incorrections.	Add.

Table 8.4: SGFL applied to the linguistic content part of the RIF for Item 1 of Activity 4 in Subtask 3 of the Learning Unit *Customer Service and International Communication* in ALLES.

8.2.2 Applying the AASF for summative assessment

In our research setting, a distinctive feature of tasks requiring summative assessment is that in producing the assessment CAF measures are used as part of the criteria for correctness in the RIF. As we described in Section 6.2.6.2, summative assessment in our context includes the assessment of communicative contents, lexical contents, sentence structure and accuracy, and overall layout. In this section we exemplify how the AASF can be applied to define the NLP specifications not only for the thematic and linguistic content, but also for the CAF measures. We will exemplify it for Activities 1 and 2 in the Final Task in the learning unit *Education and Training* in ALLES, analysed in Section 7.2.2.3.

8.2.2.1 The SALA applied to a task for summative feedback

The application of the SALA for specifications of the thematic and linguistic contents for the above mentioned task is shown in Tables 8.5 and 8.6. The specification of the language analysis module for the CAF measures is presented in Table 8.7.

The first block of specifications in Table 8.5 is related to the ENTITIES defined in the RIF. There are linguistic cues that can be expressed in different manners and in different positions in the text, as the entity *NE:EmailAuthor*, corresponding to a reference to the email's author in the text. There should be several instances of the personal pronoun *I*, and probably at least two instances of his or her name, in the introduction and in the signature. Other entities specified in the RIF would be added here, as indicated by the last row in this block.

Table 8.5 shows also the specifications derived from the RELATIONS defined in the RIF. For instance, the first row describes the NLP specifications for the criterion that evaluates whether the learner stated the department she is working in correctly – here a cross-relation with the entity *NE:DeptName* will be established in the feedback logics. For all the references to relations, the linguistic cues, the analysis codes and the NLP modules or functionalities are specified.

Reference	Textual cue	Analysis code	NLP mod.
Entities			
The email author	- *I* - *(My name is) X* - *(Sincerely,) X*	NE:EmailAuthor	IE (ent.)
Department's name	- *Marketing Department*	NE:DeptName	IE (ent.)
Course names	- *Business Communication* - *E-commerce and E-business*	NE:CourseName	IE (ent.)
(...)			
Relations			
State your department	- *I work for the Department of X* - *I am in the Department of X*	Rel:YourDepartment Rel:YourDepartment	IE (rel.) IE (rel.)
Tell what course you want to attend to	- *I am interested in the course on X* (...)	Rel:DesiredCourse	IE (rel.)
Tell why this course can benefit your in the future	- *this course will be useful for the projects I will be involved in the future* (...)	Rel:UsefulFutProjects	IE (rel.)
Tell that the course timetable suits your calendar	- *the course timetable fits with my schedule*	Rel:FitsMySchedule	IE (rel.)
Tell who authorised you to attend it	- *(and) I go the authorisation from X* (...)	Rel:PermFromDavid	IE (rel.)
Tell what courses would you be interested in	- *in the future I will be interested in X* (...)	Rel:CoursesOfInterest	IE (rel.)

Table 8.5: SALA applied to the analysis requirements for the thematic contents in tasks requiring summative assessment.

Table 8.6 shows the specification of the linguistic analyses requirements of the linguistic contents in the RIF of the task under question (see Section 7.2.2.3). The table contains different blocks for each of the description levels defined in the RIF.

As for functional content, we exemplify the exponents for two of the functions expected in the response: expressing interest and introducing oneself. Both of them require of specific IE rules in the corresponding NLP module. As for syntactic content, we exemplify the need to detect the use of verb forms in the present tense as an indicator of learners being able to describe or report states of affairs – for which a parser or a morphosyntactic tagger should be used. Functional content and syntactic content specifications should be fully completed along these lines.

As for lexical content, the RIF criteria would be converted into the requirement of a series of entries in the lexicon, or, if they are already in it, they would be converted into adding some internal code marking the relevant words as domain-specific vocabulary. As for pragmatic content, we would require IE rules to analyse the standard parts of professional emails such as a greeting, a complimentary close and a signature. This SALA block would include a list of discourse markers, which would be to a large extent lexicon work – maybe a task-specific lexicon.

The specifications in relation to CAF measures are presented in Table 8.7. The first column contains the indicators to be looked at, which correspond to quantitative measures based on linguistic and textual elements. Each of them produces a variable, in the third column, that receives a numerical value, obtained by operating on the analysis provided by the NLP modules, in the fourth column.

In the block Lexical contents, tokenisation is required to identify and count words, as an indicator for fluency – spell checking, optional, ensures that only words (not non-words) are counted. As for specific vocabulary, we would use the words marked as specific vocabulary in the lexical contents part of Table 8.6, but ensuring they are concept words, that is, adjective, nouns or verbs. Therefore, both a domain specific lexicon and a lemmatiser including POS tags are needed.

In the block Sentence structure and accuracy, there are the sub-blocks syntactic complexity and accuracy. The former is correlated with evidence from sentence types and presence of discourse markers. For this, the NLP functions identified are sentence segmentation, syntactic analysis, POS tagging and a discourse marker lexicon (or an extension of the lexicon with discourse marker information). The latter requires grammar checking, already specified in Table 8.6.

Finally the block Overall text layout, contains two sub-blocks: fluency and structure, and formal correctness. These two indicators are correlated with the number of paragraphs and the number of spelling errors, for which the functions required are paragraph segmentation and spell checking.

Reference	Textual cue	Analysis code	NLP mod.
Functional content			
Expressing interest	- *I am interested in* X	Func:ExpressInterest	IE (exp.)
Introduce oneself	- *my name is* X, *I am X* (...)	Func:IntrodOneself	IE (exp.)
(...)			
Syntactic content			
Describe/Report with Present Simple	- *I* work *for* X - *My manager* has *recommended* X	Syn:PresentTense	Morphosyntactic analysis
(...)			
Lexical content			
Domain-specific vocabulary	- work, register, apply, course, schedule, - manager, authorisation, permission (...)	Lex:DomainVocab	Lexicon
Pragmatic content			
Greeting expression	- Dear Sir or Madam - Dear colleagues, - To whom it may concern,	Prag:Greeting Prag:Greeting Prag:Greeting	IE (prag.) IE (prag.) IE (prag.)
Complimentary close	- *Sincerely yours*, (...)	Prag:ComplClose	IE (prag.)
Signature	- Proper name after the complimentary close	Prag:Signature	IE (prag.)
Discourse markers	- because, since, and (...)	Prag:DiscourseMarkers	Lexicon
Graphology			
Appropriate spelling	- Word spelling and punctuation	SpellingOk	Spell/Grammar checker

Table 8.6: SALA applied to the analysis requirements for the linguistic contents in tasks requiring summative assessment.

Reference	Textual cue	Analysis code	NLP mod.
Lexical contents			
Word fluency	- Length in words	NW: No. of words	Tokenisation, spell checking (opt.)
Specific vocabulary	- Domain specific terms	SV: Specific vocabulary	Lemmatisation and domain-specific lexicon
Sentence structure and accuracy			
Syntactic complexity	- Simple and complex sentences - Presence of discourse markers	NS: No. of sentences (simple and complex) NDM: No. of discourse markers	Sentence segmentation and syntactic analysis POS tagging and discourse marker lexicon
Accuracy	- Grammar errors	NGE: No. of grammatical errors	Grammar checking
Overall text layout			
Fluency and Structure	- Organisation in paragraphs	NP: No. of paragraphs	Paragraph segmentation
Formal correctness	- Spelling errors	NSE: No. of spelling errors	Spell checking

Table 8.7: SALA applied to the analysis requirements for the application of CAF measures in tasks requiring summative assessment.

8.2.2.2 The SFGL applied to an activity with summative feedback

The specification of the feedback generation logic for an activity with summative feedback requires the definition of how the feedback messages should be generated according to the hypotheses based on the automatic linguistic analysis. In Section 7.2.2.3.2, we presented a detailed characterisation of the criteria for summative assessment as provided by the RIF specifications, which can be converted into a feedback generation logic by applying the SFGL.

Table 8.8 shows the results of applying it to a selection of the assessment conditions. For simplicity, the columns priority, match and type are not included: For this particular task, priority is always high and only full matching is considered. As for the message type, all of them are considered of type Main within each block, that is, there will be one and only one message and grade for the block COMMUNICATIVE CONTENTS, another one for the block LEXICAL CONTENTS, and so on.

A difference in Table 8.8 with respect to the SFGL specifications of a task with formative feedback is that it includes a *grade* associated to each feedback message. This is the number that provides the summative assessment, a value between 0 and 4 for each of the blocks. Thus, the learner's response is assessed upon a total of 16 points, four for each of the blocks as required by the RIF specifications.

The table should be read as following: the first cell in each row is the analysis condition, where it says which are the values for each to the variables defined (in Tables 8.6 and 8.7) that yield specific grades and messages. For instance, the very first row states that if all the expected thematic content and linguistic content items are identified, that is, if $TC = 6$ and ($\wedge$) $LC = 4$, then the response receives a 4 in the assessment of communicative contents. Besides, the message "Very good. You use the expected functions adequately." is associated with the grade. The other rows and blocks contain different variables and variable names that would be combined in a similar way.

Analysis condition	Grade – Message
Communicative contents	
$TC = 6 \wedge LC = 4$	4 – Very good. You use the expected functions adequately.
$TC = 6 \wedge LC = 3$	3 – Very good. You use almost all of the expected functions adequately.
(...)	
$TC = 4 \wedge LC = 1$	0 – Are you sure you have understood the purpose of this exercise?
$TC \leq 3 \wedge LC \leq 4$	0 – Are you sure you have understood the purpose of this exercise?
Lexical contents	
$80\% \leq SV \leq 100\% \wedge 90\% \leq NW \leq 100\%$	4 – Excellent. Your text reads well and is precise. You are using the (...)
$80\% \leq SV \leq 100\% \wedge 50\% \leq NW \leq 89\%$	3 – Good. Your text is pertinent but you should be more fluent.
(...)	
$0\% \leq SV \leq 29\% \wedge 50\% \leq NW \leq 89\%$	1 – Careful; Try to be more fluent. Check the vocabulary you are using!
$0\% \leq SV \leq 29\% \wedge 0\% \leq NW \leq 49\%$	0 – Careful. Your vocabulary is inappropriate and the text does not read well.
Sentence structure and accuracy	
$10 \leq NS \leq 9 \wedge 10 \leq NDM \leq 9 \wedge 0 \leq NGE \leq 1$	4 – Great. Your text is correct and adequate. There are no mistakes.
$NS = 8 \wedge 10 \leq NDM \leq 9 \wedge 0 \leq NGE \leq 1$	4 – Great. Your text is adequate, but there are some minor errors.
(...)	
$5 \leq NS \leq 0 \wedge 6 \leq NDM \leq 10 \wedge NGE \geq 3$	1 – Careful. Some information is missing and the text has some grammatical mistakes.
$0 \leq NS \leq 7 \wedge NDM \leq 5 \wedge NGE \geq 3$	1 – Careful. Some information is missing and you are not using any connecting words.

Table 8.8: SFGL applied to the summative assessment criteria for Activities 1 and 2 in the Final Task in *Education and Training* in ALLES (continues).

Analysis condition	Grade – Message
Overall text layout	
$NP = 9 \wedge 0 \leq NSE \leq 1$	4 – Excellent. Your text has an adequate structure and no spelling mistakes.
$NP = 9 \wedge 2 \leq NSE \leq 3$	2 – Good. Your text has an adequate structure, but some spelling mistakes.
(...)	
$NP \leq 5 \wedge 2 \leq NSE \leq 3$	1 – Your text does not have any structure. There are some spelling mistakes.
$NP \leq 5 \wedge NSE \geq 4$	0 – Careful: Your text does not have any structure and has many spelling errors.

Table 8.8: Application of the AASF to the Activities 1 and 2 in the Final Task in *Education and Training* in ALLES.

8.3 A feedback generation strategy for the assessment of ICALL activities

A pedagogically informed NLP-based automatic assessment module can be implemented on the basis of SALA and SFGL specifications. The task at hand is to convert those specifications into a feedback generation strategy to provide learners with automatic feedback.

Figure 8.3 reflects the system-learner interaction by detailing the role of the Automatic Assessment Module. In the Linguistic Analysis module (LA), learner responses are automatically analysed by NLP tools, a process through which a linguistically annotated version of the learner response is obtained. On the basis of this automatic annotation, the Feedback Generation module (FG) is responsible for evaluating the performance of the learner by relating the analysed response and the modelled assessment criteria. The figure suggests an iterative process, a key characteristic for learning materials for which formative assessment is foreseen.

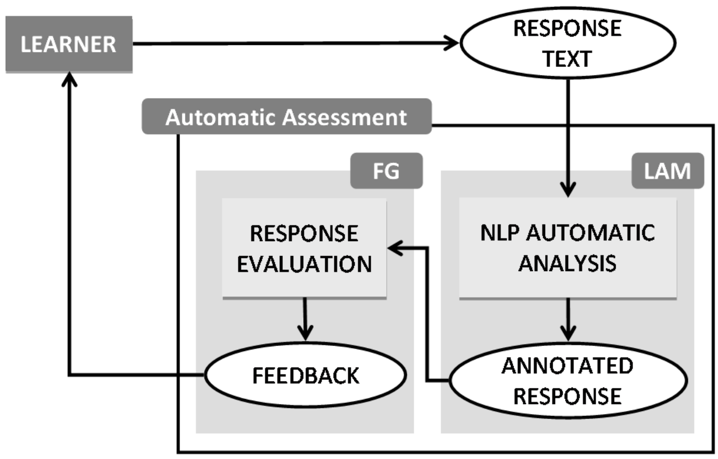


Figure 8.3: System-learner interaction for the evaluation of responses with the Automatic Assessment Module.

Figure 8.4 presents the actual implementation of an Automatic Assessment module for all the tasks in our research setting (Boullosa, Quixal, Schmidt, Esteban, and Gil, 2005: pp. 20). This first step aims to reduce the number of formal errors, so that when the response is analysed with the task-specific modules for the analysis of more complex linguistic structures an optimal performance can be ensured. This initial correction process is a strategy to reduce the effects of ill-formed language on the the performance of NLP tools. Since the accuracy of the domain-independent spell and grammar checkers is also imperfect (false positives), learners are given the chance to ignore it altogether and proceed to the second correction step.

This two-step procedure is also related to the use of two different types of resources. On one hand, domain-independent NLP resources are used in the first correction step – the resources described in Chapter 6. On the other hand, there are the domain-dependent resources, which are responsible for the analysis and evaluation of the response given the pedagogical setup of the task.

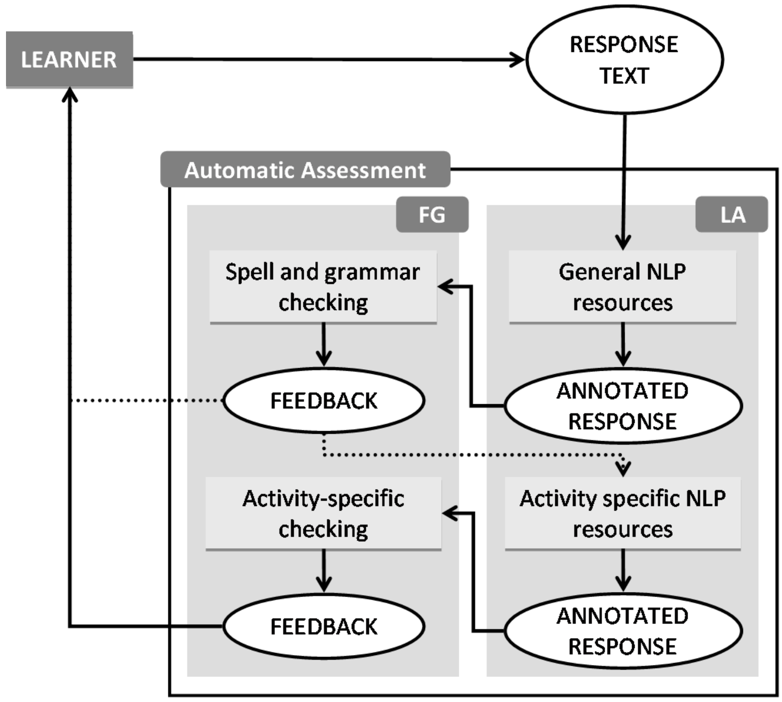


Figure 8.4: Two-step feedback presentation flux in ALLES.

8.3.1 A general NLP-based architecture for the automatic assessment of learner responses

In this section we present a generalised feedback generation architecture based on the one presented in Chapter 6. Figure 8.5, as opposed to Figure 6.2 in Section 6.3, includes a module called Global Response Checker, responsible for the actual generation of feedback on the basis of the SFGL specifications. This feedback generation architecture foresees a customisation of part of the Linguistic Analysis modules (LA) and the Feedback Generation modules (FG) for each ICALL task following the respective SALA and SFGL specifications. The details of the conversion of SALA/SFGL specifications into rules are described in the following section.

8.3.2 The point of departure for NLP-based automatic assessment

As shown in Figure 8.5, the morphosyntactic level of analysis is the point of departure for task-specific analysis – indicated by a dotted arrow from the module Morphological Disambiguator to the Information Extraction Modules. Thus, before we start describing how more complex levels of linguistic analysis can be implemented we present Table 8.9, which contains a representation of a sentence analysis at the levels of token, lemma and morphosyntactic features. The sentence for which the analysis is represented is *How satisfied are you with Stanley Broadband?*, one of the possible correct answers to Item 1 of Activity 4 in Subtask 3 of the Learning Unit *Customer Service and International Communication* (see Section 7.2.2.1).

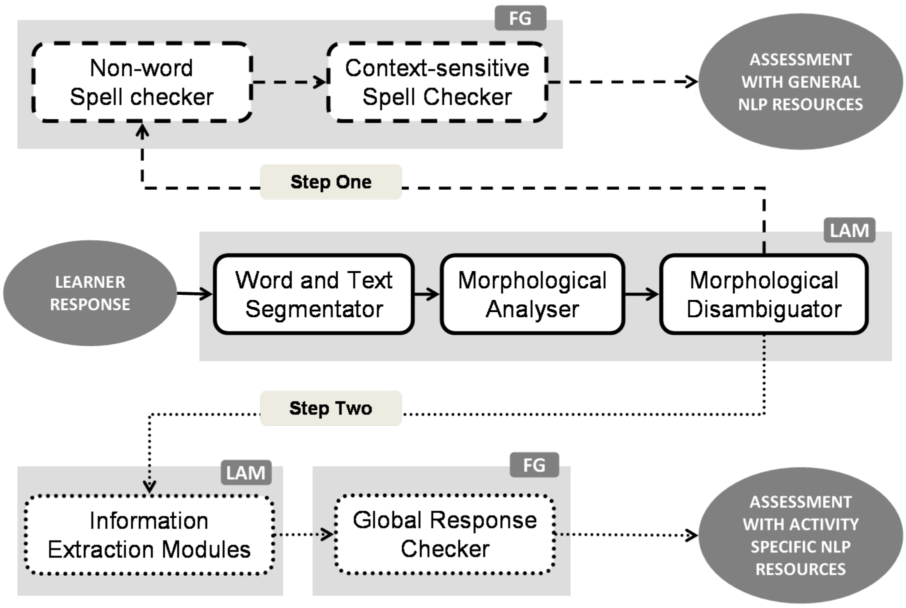


Figure 8.5: A domain-adaptive NLP-based feedback generation architecture for formative and summative assessment. Dotted lines indicate domain-specific resources.

The first column in Table 8.9 is a token identification number.[1] The second column shows the result of segmenting the sentence into tokens, and identifying the sentence boundaries, as a result of applying the tokenisation module. The third and the fourth column show the result of lemmatisation and morphosyntactic disambiguation. The result contains a lemma, a non-inflected version of the word, a grammatical category (adverb, verbs, pronoun, noun, preposition or, a textual category, punctuation), and associated morphosyntactic information to some of the words: nouns have a gender and number associated, verbs have tense, mode, person and number, and so on.

ID	Token	Lemma	Morph. information
	<s>		
1	*How*	how	adv wh-word
2	*satisfied*	satisfy	verb part
3	*are*	be	verb pres ind 2nd pers pl-sg
4	*you*	you	pron 2pers pl-sg
5	*with*	with	prep
6	*Stanley Broadband*	Stanley Broadband	noun sg proper
7	*?*	?	punct
	< /s>		

Table 8.9: Abstract representation of the linguistic analysis obtained with the general module of the architecture.

Note also that Table 8.9 shows the sequence of words *Stanley Broadband*, token no. 6, is as a proper noun – assuming there is a general heuristics for named entity recognition. The techniques applied for entity recognition could determine the order in the linguistic processing but this is not relevant for our research purposes.

[1]The notation in table does not reflect the actual data structure. See footnote [4] in page 94.

8.4 Automatic generation of formative feedback

In this section we describe the implementation of the rules underlying the feedback generation strategy for the provision of formative feedback, both for Information Extraction modules and the Feedback Generation module. The strategy combines domain-specific, that is, task-specific, linguistic analysis modules and feedback generation modules implemented using finite-state techniques. On the basis of morphosyntactic and lexical information, Information Extraction techniques will help us obtain the analyses to judge the correctness and well-formedness of learner responses.

All through this section we will use Item 1 of Activity 4 in Subtask 3 of the Learning Unit *Customer Service and International Communication* as the sample response to be modelled – its RIF details are given in Section 7.2.2.1, Table 7.7.

8.4.1 Modelling automatic assessment for correct responses

The sentences in (10) are a list of RIF-based correct responses to Item 1 of Activity 4 in Subtask 3 of the Learning Unit *Customer Service and International Communication*. We will use this set of responses to exemplify the implementation of the linguistic analysis resources that allow for the handling of correct responses.

(10) a. How satisfied are you with Stanley Broadband?
b. How happy are you with Stanley Broadband?
c. How much do you enjoy Stanley Broadband?

Assuming the linguistic analysis shown in Table 8.9 for (10a) is available for all of the above sentences, the extraction of the expected information using rule-based strategies can be implemented for all task-specific characteristics beyond morphosyntactic tagging. This is described in the following sections.

8.4.1.1 Modelling linguistic analysis of correct responses

The graph in Figure 8.6 shows an abstraction of a finite-state automaton strategy to analysed the thematic and linguistic contents of the possible correct response in (10). Complex linguistic structures can be analysed on the basis of lexico-mophosyntactic patterns, whose number of elements can range from one to six or seven. The resulting complex linguistic structures will be coded according to the SALA specifications.

Thus, the initial node in the graph allows for the detection of the property *Synt:HowFirst*, which can be later correlated with the criterion that requires that the response starts with *How*. This would be a rule in a IE module specialised on the analysis of the syntactic contents of the RIF.

The edges that go out of the first node offer two alternative paths to be followed, and both of them identify the relation *Rel:LevelOfCustomerSatisfactionWith*, as specified in the thematic contents of the RIF. These two paths correspond to two different linguistic constructions. One of them uses the auxiliary *do* and the verb *enjoy*, a lexical choice that requires a direct object as a complement, a function that can be performed by a noun phrase, such as *Stanley Broadband*, or a subordinate clause as the one started by *using*. Note that the optionality of *using* is marked by two possible internal paths, one that allows a transition from the node *enjoy* directly to the node *Stanley Broadband* and the other one that goes first through the node *using*. The other construction requires the copulative structures *be* + ADJECTIVE or *be* + PAST PARTICIPLE, both of which have to be followed by *with*, the preposition governed by *happy* and *satisfied*.

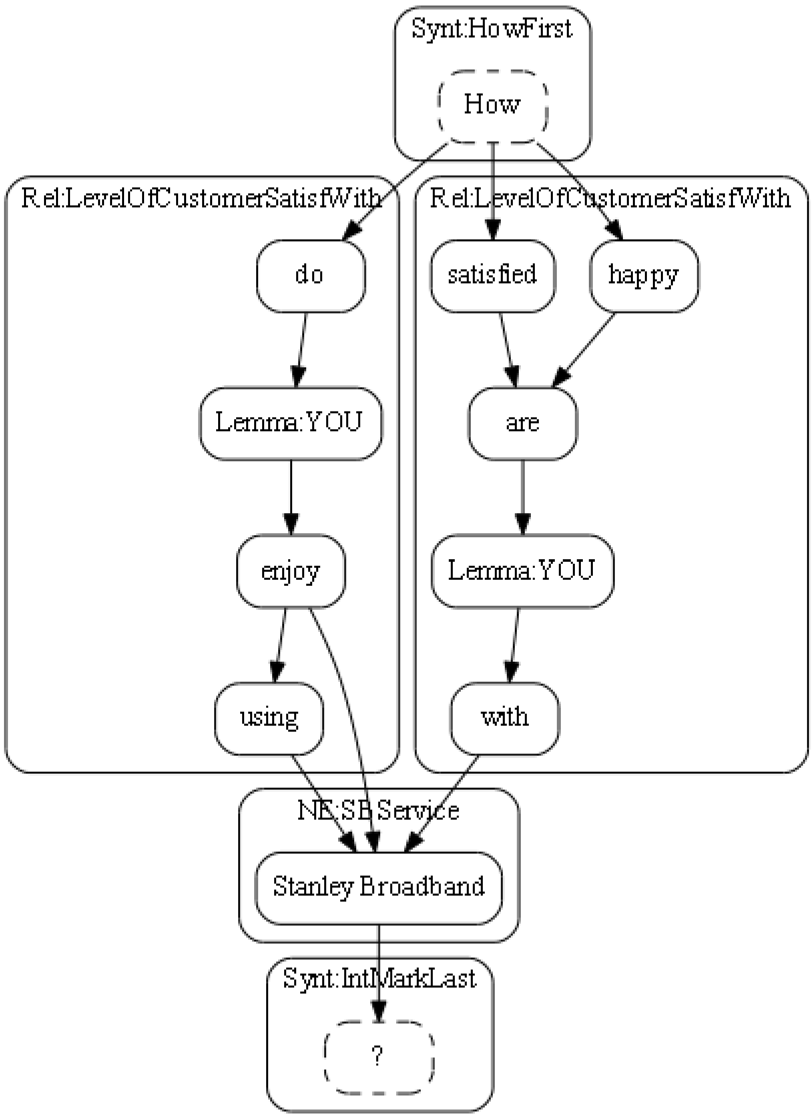


Figure 8.6: Recognition paths to analyse linguistic structures relevant for the assessment of thematic and linguistic contents of the responses in (10).

The two last single token nodes in Figure 8.6 correspond to linguistic elements that can be identified by the morphosyntactic tagging analysis. However, for the Feedback Generation module to properly evaluate the correctness of the response task-specific description labels are assigned. If all the complex linguistic elements (*Synt:HowFirst*, *Rel:LevelOfCustomerSatisfactionWith*, and so on) are correctly identified, the FG module will be able evaluate the correctness of the response.

8.4.1.2 Modelling feedback generation of correct responses

The criteria for correctness to be checked for after the linguistic analysis are implemented using FSA techniques that search for sequences of relevant analysis codes. Figure 8.7 shows the recognition path required to check for the correctness of the response to the task's item in question according to the criterion, the analysis condition.

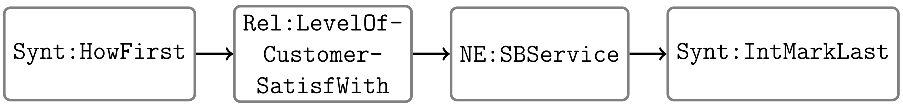


Figure 8.7: Global response evaluation recognition path for the response to Item 1 in the customer-satisfaction-questionnaire activity.

If the linguistic analysis of a response allows for crossing the nodes in FSA in Figure 8.7, the response will be considered correct and the message for the full matching condition will be triggered.

8.4.1.2.1 Response Global Checker in KURD

In order to provide a link with the technical description provided in Section 6.3.2.1.1, we describe the KURD rules that would be included in the Response Global Checker to check for this criterion. The rule in Figure 8.8 checks for the presence and the correct sequencing of the response contentsfor the task item. If the condition applies, an *assessment* feature is added to the sentence analysis node with the information *correct* at the *global* response level – the relevant line of code is line no. 15.

```
2  CustomerSatisf =
3          Ae{ori=How},
4          Ae{ori=satisfied},
5          Ae{ori=are},
6          Ae{ori=you}
7       :
8          $-1g{disc=CustSatisf},
9          Ar{disc=CustSatisf},
10         j(rule=@Prod).
11
12 CorrectResponse =
13     ?$-1a{RespOrder=a_HowDegOne_CustSatisf_Prod_HowDegTwo}
14     :
15     $-1g{assessment={type=global,code=correct}},
16     j(rule=NEXT).
```

Figure 8.8: KURD rules to process a part of a possible response to one of the ICALL activities later on presented and worked out in Chapters 7 and 8.

8.4.2 Modelling incorrect responses

In addition to modelling correct responses, our feedback generation strategy requires the modelling of incorrect responses to be able to inform learners about possible errors. Using a shallow parsing strategy based on FSA techniques for this requires the ability to anticipate different degrees of deviating structures – derived from or related to the set of gold-standard responses.

For the purpose of this explanation, we have made up examples that help us exemplify the different techniques and levels of linguistic analysis that can enhance NLP-based feedback generation systems to better handle incorrect responses.

8.4.2.1 Modelling wrong choices in responses

We will present modelling strategies for the detection of errors based on three different types of transformations of the expected response: wrong choices, that is, use of a word or expression in a response as a substitution of a correct one, and missing or unexpected information.

8.4.2.1.1 Linguistic analysis of responses with wrong word form errors

By using levels of linguistic analysis more complex than the word level we can assess sentences such as the ones presented in (11), all of which qualify as wrong choice errors.[2]

(11) a. How *satisfying* are you with Stanley Broadband?
b. How satisfied *is* you with Stanley Broadband?

To better understand how this strategy is implemented, let us first compare the analysis of the correct version of the response, namely *How satisfied are you with Stanley Broadband?*, with the analysis of the incorrect response, the sentence in (11a). In Table 8.10 a small box highlights the linguistic features that differ between the correct response and the incorrect one in terms of lemmata and the grammatical information. The deviating response contains the present participle verb form, not the past participle one, of the correct lexical choice. Thus the matching would succeed at the lemmata level.

Correct response		Incorrect response	
Lemma	Grammatical info.	Lemma	Grammatical info.
how	pron	how	pron
satisfy	verb part [past]	satisfy	verb part [pres]
be	verb pres ind 2pers sg	be	verb pres ind 2pers sg
you	pron 2pers sg	you	pron 2pers sg
(...)		(...)	

Table 8.10: Comparison of the analysis of Example (11a) with the one of the correct response *How satisfying are you with StanleyBroadband?*

[2]Along this and the following sections we use deviating sentences based on the correct response *How satisfied are you with Stanley Broadband*?, but the explanation and techniques are equally applicable to any of the alternative correct responses seen in the previous section.

Figure 8.9 shows a recognition path, based on the one shown in Figure 8.6, that exploits lemmata information to analyse the linguistic structures in the response. The difference between two automata is that the one in Figure 8.9 contains nodes where the element to be recognised is a lemma: marked with the prefix “Lemma:” and a non-inflected version of the word.

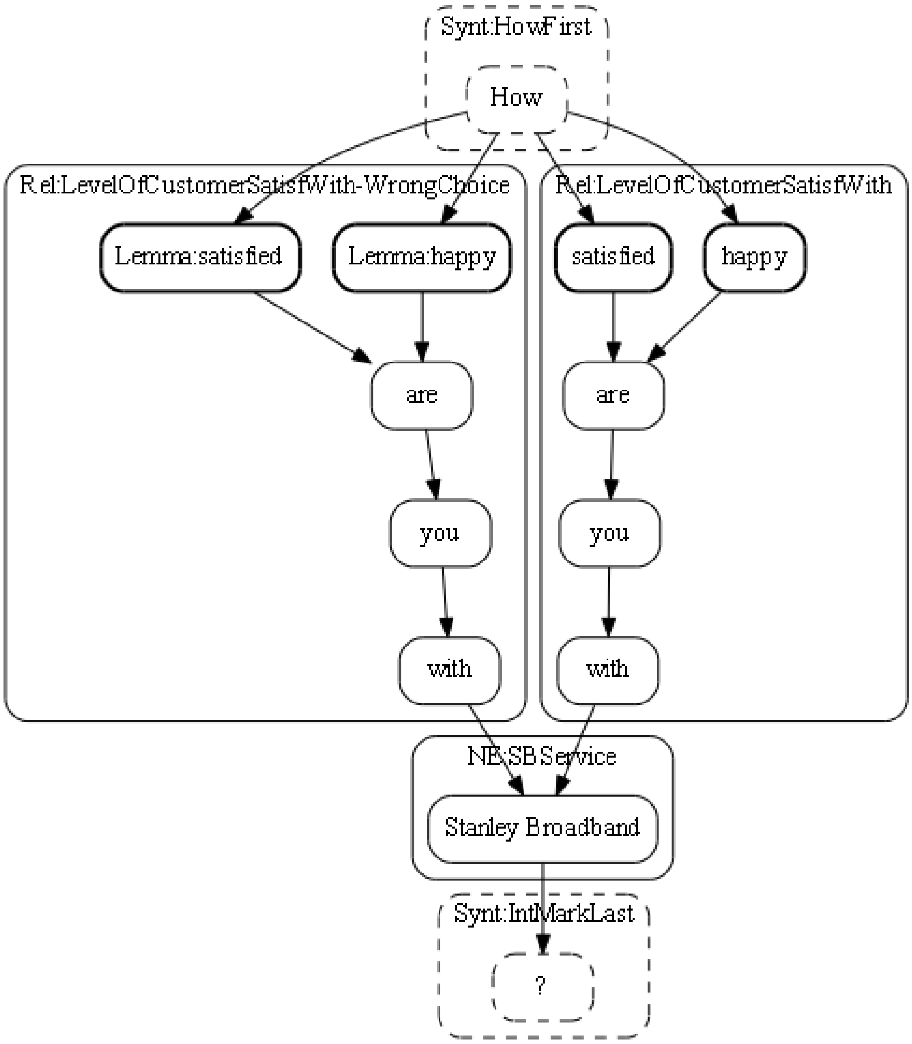


Figure 8.9: Recognition path to analyse responses including wrong choice errors as the one in Example (11a).

With this strategy, the response can be detected as correct in terms of thematic contents though it is not correct, or not as expected, in terms of linguistic form. In Figure 8.9 this is marked by the adding the suffix *WrongChoice* to the analysis code *Rel:LevelOfCustomerSatisfactionWith*, the one we used for the analysis of this relation the modelling of correct responses.

Table 8.11 compares the analysis of the incorrect response in (11b), *How satisfied is you with Stanley Broadband?*, with the correct one. In this case, the deviating response contains the present third-person singular form of the verb *to be*, not the present second-person singular form. The matching may succeed at the lemmata level, with the caution to tag the deviations properly so that the FG module can then be informed.

Correct response		Incorrect response	
Lemma	Morphosynt.. analysis	Lemma	Morphosynt.. analysis
how	pron	how	pron
satisfy	verb part past	satisfy	verb part past
be	verb pres ind 2pers sg	be	verb pres ind 3pers sg
you	pron 2pers sg	you	pron 2pers sg
(...)		(...)	

Table 8.11: Comparison of the analysis of Example (11b) with the one of the correct response *How satisfying are you with StanleyBroadband?*

Figure 8.10 shows the two possible recognition paths for the relation *Rel:Level-OfCustomerSatisfactionWith*, one of which is marked with the suffix *WrongChoice*.

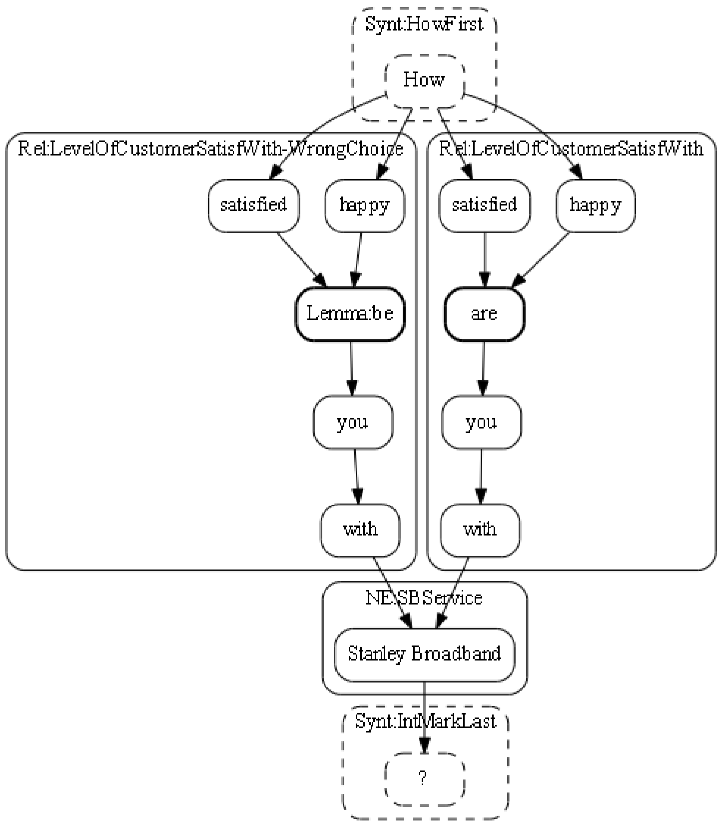


Figure 8.10: Recognition path to analyse a response including wrong choice errors as the one in Example (11b).

8.4.2.1.2 Linguistic analysis of responses with wrong lexical choice errors

Another type of wrong choice error to be handled is shown in (12), where an incorrect preposition is used. The construction BE SATISFIED + PREP collocates with the preposition *with* – or *by* if we consider a passive construction –, but not with *against*.

(12) How satisfied are you *against* Stanley Broadband?

The problem is the lexical choice, the "word" in itself is wrong, but the fact that a preposition is used in its place allows us to use morphosyntactic information to analyse and diagnose the problem.

Table 8.12 reflects the relevant differences in the linguistic annotation generated by the automatic analysis for each of the correct and the incorrect version of the sentence. Exploiting the information at the level of grammatical category and adding the condition that the relation *Rel:LevelOfCustomerSatisfactionWith* can also be identified if a preposition with a different lemma is used would allow for a successful recognition of this structure.

CORRECT RESPONSE		INCORRECT RESPONSE	
LEMMA	MORPHOSYNT.. ANALYSIS	LEMMA	MORPHOSYNT.. ANALYSIS
how	pron	how	pron
satisfy	verb part past	satisfy	verb part past
be	verb pres ind 2pers sg	be	verb pres ind 2pers sg
you	pron 2pers sg	you	pron 2pers sg
with	Prep	against	Prep
(...)		(...)	

Table 8.12: Comparison of the correct response with the analysis of *How satisfied are you* against *Stanley Broadband?*

Figure 8.11 shows the automaton that recognises the incorrect version of the sentence, the one including *against*. It contains a node that accepts a word with a preposition reading whose lemma is not *with*: Prep LemmaNOT:with.

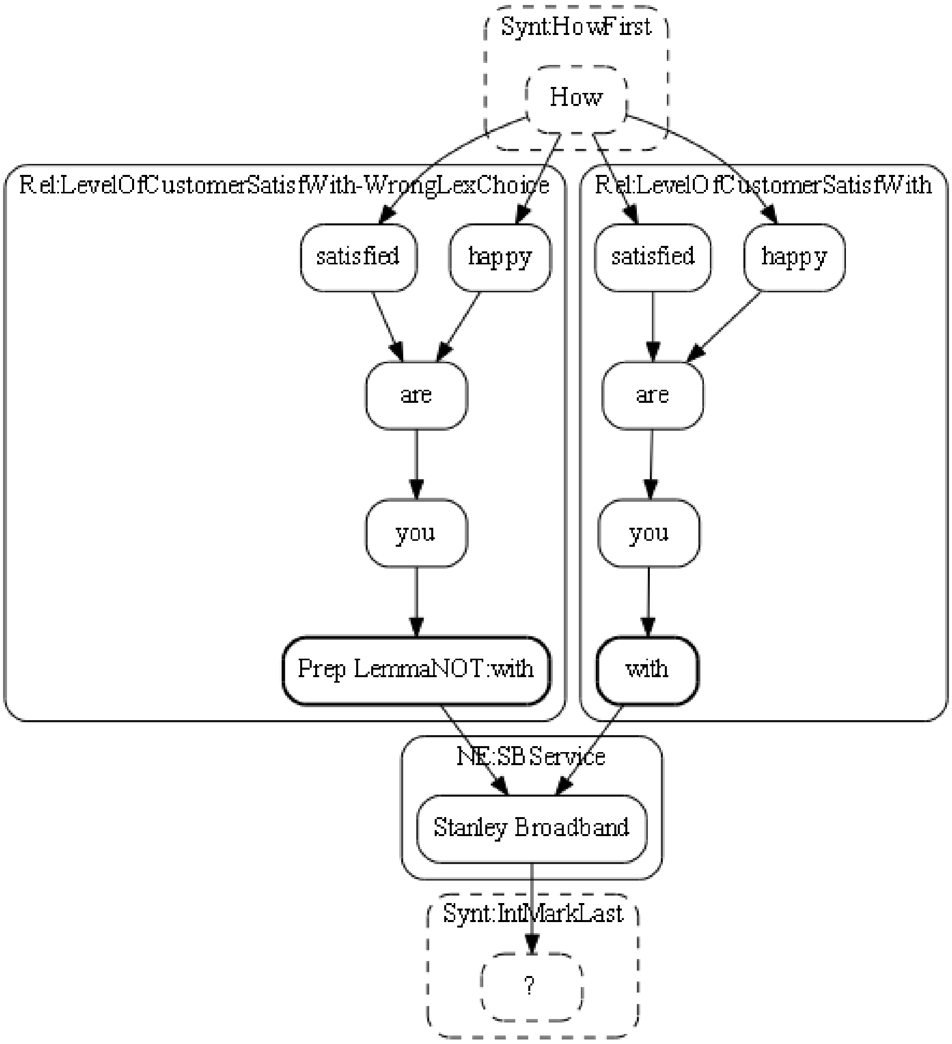


Figure 8.11: Recognition path to analyse a response including wrong choice errors as in Example (12).

8.4.2.1.3 Feedback generation for responses with wrong choice errors

The implementation of mal-rules for the linguistic analysis is accompanied by a strategy to generate the corresponding feedback. In this sense, a response evaluation automata used for the evaluation of correct responses as the one presented in Section 8.4.1.2 can also be applied for the evaluation of incorrect responses. However, there will be an algorithm responsible for comparing the differences between the correct and the incorrect responses. Such an algorithm will look for suffixes as the ones we described (-*WrongChoice*) and compute the feedback messages with what such analyses would be correlated.

Typically, if the deviation is related with a wrong form the feedback strategy, the assumption is the response is correct in terms of thematic contents, but not in terms of linguistic contents. If the deviation is related to a wrong lexical choice, and particularly with grammatical categories that tend to have a semantic weight in the sentence (e.g., nouns, verbs, and adjectives), then the feedback strategy will assume the response is incorrect in terms of thematic contents.

8.4.2.2 Modelling missing or unexpected information

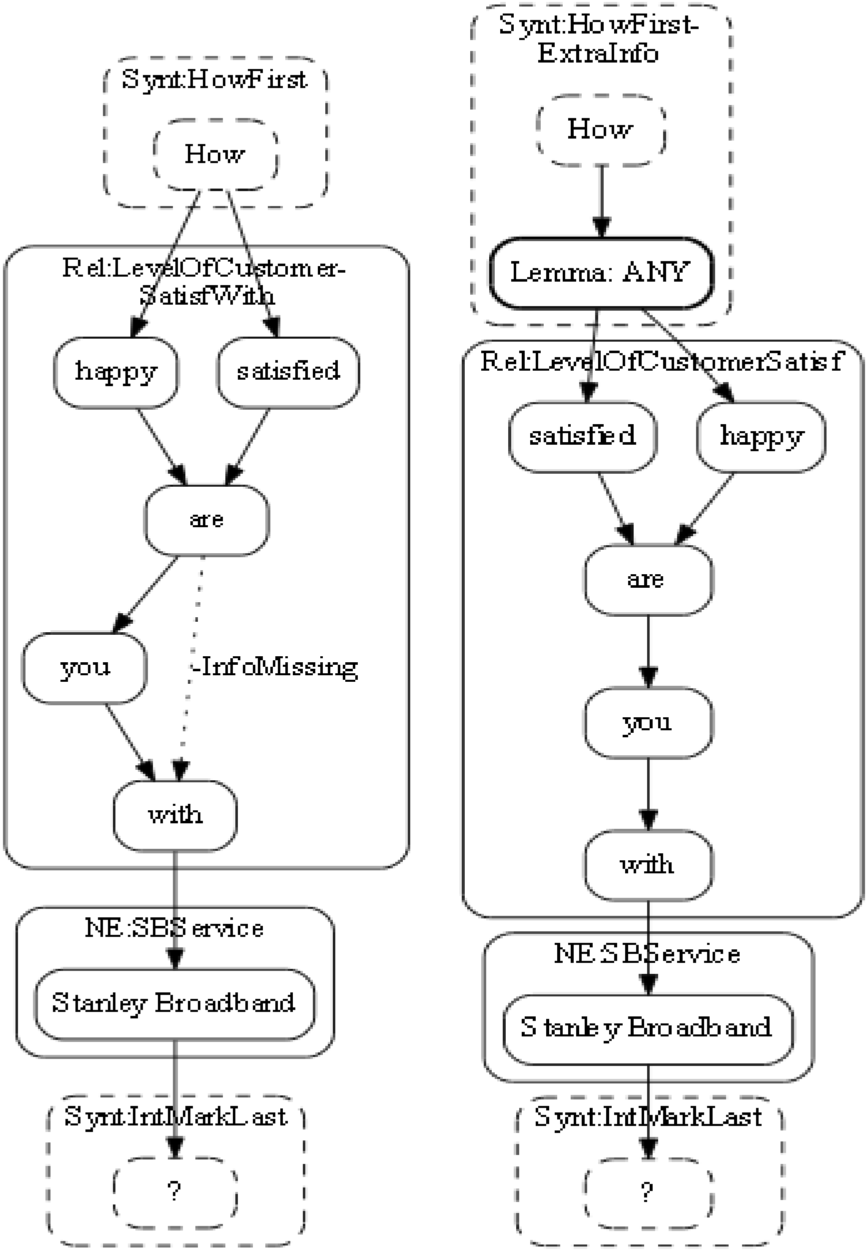
The sentences in (13) are examples of responses providing more or less information than expected. In (13a) the pronoun *you* is missing, marked with the symbol of the empty set $\emptyset$; in (13b) the word *much* was added between *How* and *satisfied*.

(13) a. How satisfied are $\emptyset$ with Stanley Broadband?
b. How *much* satisfied are you with Stanley Broadband?

Linguistic modelling of responses with missing or extra information The modelling of these two types of deviations requires recognition paths including specific nodes to foresee the presence or absence of specific information. Figure 8.12 shows the recognition paths needed to process sentences as the ones in (13).

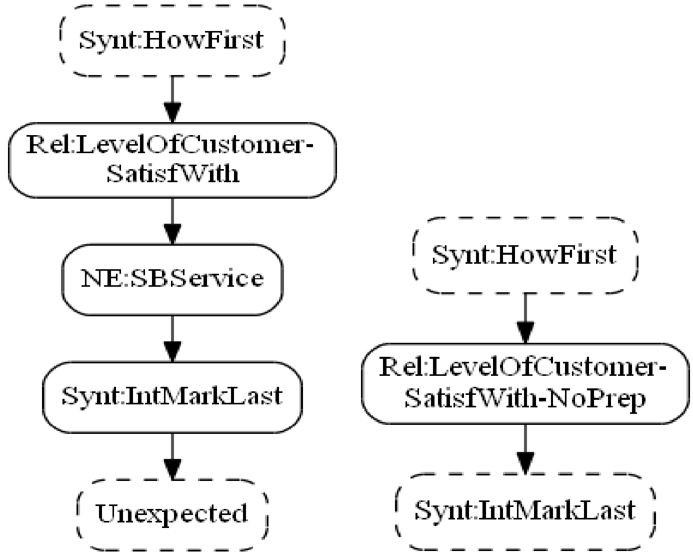
Figure 8.12a shows two alternative paths after the word *are*. One of them requires the word *you*, the other one allows to cross from the node *are* to the node *with* by adding the label *-MissingInfo* to the analysis code as the label in the dotted arrow shows. Figure 8.12b shows an initial node allowing to parse the response element *Synt:HowFirst* including an expected element in the second position, and adding the suffix *-ExtraInfo*.

Feedback generation for responses with missing or extra information The adapted analysis recognition codes for ill-formed text can also be used by the response evaluation automata used for correct responses. These analysis codes will generate partial matchings of the analysis conditions as defined in Tables 8.3 and 8.4, and an internal algorithm will be responsible for generating the feedback.



(a) Pattern with missing elements. (b) Pattern with extra elements.

Figure 8.12: Graphs representing the recognition paths for modelling response components with missing or unexpected information.



(a) Patterns with extra elements

(b) Patterns with missing elements

Figure 8.13: Graphs representing the recognition paths for global response evaluation of responses with missing or unexpected response components.

8.4.2.2.1 Missing or unexpected information at the level of global response

Yet another different type of variation related to information that is missing or unnecessary is the one that affects the global response, that is, the number of information chunks as expected, as in the sentences in (14).

(14) a. How satisfied are you with Stanley Broadband? *And how unsatisfied?*
 b. How satisfied are you ∅?

In (14a) the response contains the text *And how unsatisfied?*, an addition to the text that, without this, would suffice for a correct response. As for the sentence in (14b), it contains a response where the reference to the service, *Stanley Broadband*, is missing, as well as the preposition *with*.

Figure 8.13 reflects the kind of modelling for these types of deviations. Note these are the paths for the evaluation of the analysis conditions in the feedback generation module. Internal nodes determining the paths for linguistic analysis are presupposed.

For the sentence in (14a), the evaluation rule presupposes a "dummy" node for the recognition of sequences that contain words that appear after the expected elements of the response and/or words that are not expected in the response. All these would be mapped into an analysis code *Unexpected*.

For the linguistic analysis of the sentence in (14b), we would use a strategy similar to the one used for the example in which *you* was missing. This is reflected in Figure 8.13b, where the *NE:SBService* node is not required, and the suffix -*With-NoPrep* is added to the node *Rel:LevelOfCustomerSatisfaction*.

8.4.3 Modelling extended production responses

In this section we want to pay attention to some key aspects of the modelling of the linguistic analysis module and the feedback generation logics for activities requiring extended production responses. These considerations refer to activities requiring longer responses, as the ones analysed in Sections 7.2.2.2, and 7.2.2.3.

We will take as example the FL learning task analysed in Section 7.2.2.3, in which learners are expected to write an email to register for a course. We first present the implementation of the evaluation rules of the Global Response Checker, because it will be easier to go from the more general elements of the response to more concrete.

If we recall the analysis of the response in terms of criteria for correctness (Section 7.2.2.3.3), the text to be produced has to consist of the ten information chunks as reflected in Figure 8.14.

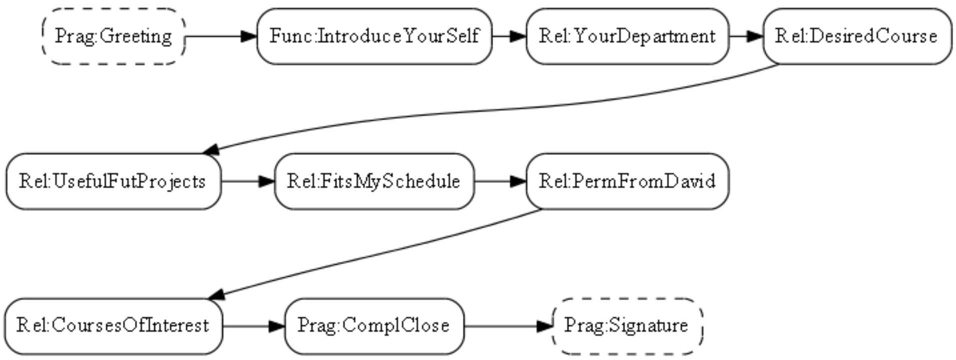


Figure 8.14: Recognition path of the response components of a language learning activity with an extended production response.

It is an email that has to comply with the formal requirements of the text type: greeting, body, complimentary close and signature; and the informative requirements of the communicative setting: introduce yourself and your department, state the course you want to attend and that the timetable and your schedule fit, name the person who approved your attendance, argue how it will be useful for you, and express interest in other courses in the future.

Figure 8.14 is an implementation of the SFGL at the level of global response, but for each of the nodes in this automaton, a corresponding set of automata would be required in the Information Extraction Modules – so that the corresponding linguistic structures are properly analysed and labelled.

Figure 8.15 shows the recognition path for the response element “Course”, which is the one in which the learner is expected to communicate for which course is he or she willing to register. We expect to find a piece of language similar to “I want to register for the course(s) on X and Y.”

Note that Figure 8.15 reflects optional paths in different ways. For instance, a transition from am to interested can succeed through very or directly. This reflects different possible syntactic structures of expressing the idea to be contained in the response component. Optionality at the lexical level is marked with disjunctions within the node. For instance, we can go from am to to through planning or

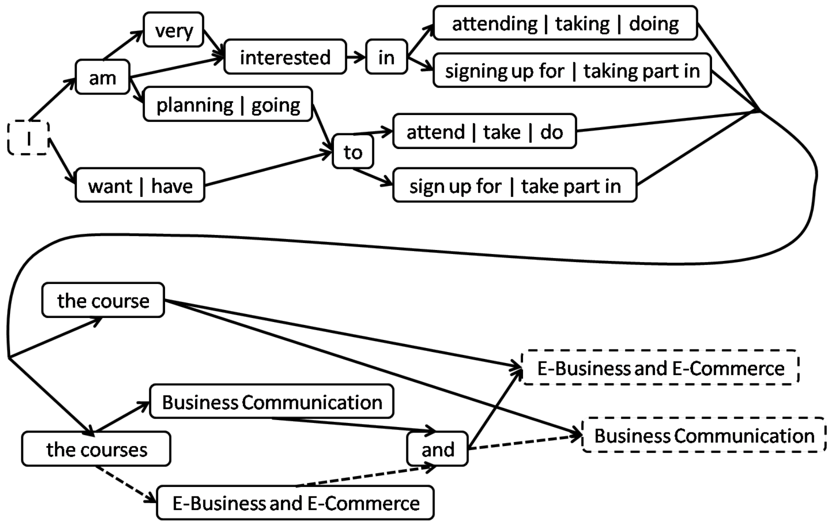


Figure 8.15: Partial recognition path of the element "Course and availability" for the response to the Final Task of the learning unit *Education and Training*.

going, and we mark this using a vertical bar (|) between *planning* and *going*, both in the same node.

If we extrapolate from the explanations of the modelling strategies used for the analysis and evaluation of correct and incorrect responses to limited production responses – Sections 8.4.1 and 8.4.2, we can think of the need to design recognition paths other than the ones shown in Figures 8.14 and 8.15 to be able to handle a variety of responses within the limits of the RIF specifications.

A critical question, though, will be the balance between flexible recognition paths, so that minor differences do not hinge the response elements to be identified, and analysis strategies that ensure a minimum linguistic structure in the response. For instance, a recognition path as the one in Figure 8.16b models responses in a very relaxed fashion. In this case the automaton would only care for detecting sequences such as *course (...) Business Communication*, *course (...) E-Business and E-Commerce*, and so on. The drawback of this strategy is that it ignores the formal linguistic aspects of the response and then the system is not able to monitor whether there is evidence of the language knowledge that learners show.

To compensate such flexible recognition paths one can add specific heuristics (rules) in the domain-specific analysis or feedback generation modules that check for specific structures. For instance, in the formal analysis we performed for the register-for-a-course activity in Table 7.11 in Section 7.2.2.3, it was determined that for that particular activity it was relevant to be able to express ones own interests and intentions. If we consider that it is important that the learner uses expressions such as *interested in* + VERB -ING, *would like to* + INFINITIVE, etc. then these heuristics should be developed so that deviating variations of this type of linguistic structure are identified. To do this, we would use mal rules whose implementation would resemble the graphs we used in Figures 8.10 and 8.11 to represent the recognition of deviations with respect to form but not lemma (recall the *satisfying* vs. *satisfied* variation) or with respect to lexical choice but not grammatical category (recall the

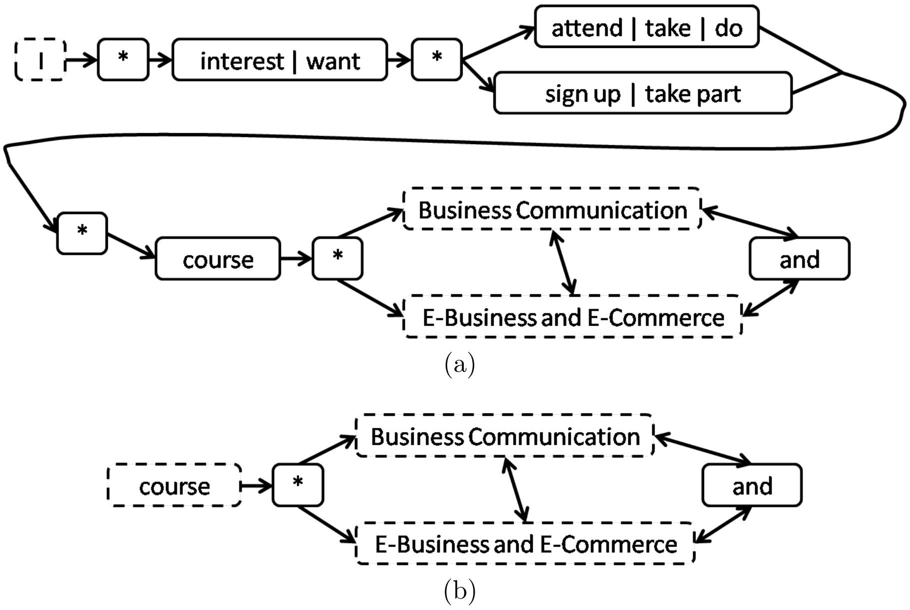


Figure 8.16: Flexible recognition paths using a “bag-of-words” approach to response component analysis.

against vs. *with* variation). However, this would require a complexity and efforts that would have to be evaluated and ideally contrasted with empirical evidence of actual learner behaviour for this kind of task.

Response Global Checker in Constraint Grammar

As we did before for KURD, we make now a short parenthesis to show how the CG rules for the Global Response Checker look like. Once the corresponding information is annotated by the Information Extraction module, a rule like the one in Figure 8.17 can be applied. The rule corresponds to the checking facilities for the Catalan version of the course-registration task. It checks for the presence of all the response elements expected for this language learning activity and checks also whether a response component is missing. If so a corresponding code is mapped onto the text asa whole.

```
ADD (@:ComplCloseMissing) TARGET (EndOfText) IF
                    (0 Greeting) (*1 EmailReason
                    LINK *1 PlaceConfirmation BARRIER ResponseLimit
                    LINK *1 IncludedAttachment BARRIER ResponseLimit
                    LINK *1 Thanking BARRIER ResponseLimit
                    LINK *1 NOT ComplClose BARRIER ResponseLimit);
```

Figure 8.17: CG-based rule in the Global Response Checker to be applied to the Catalan version of the activity described in section 7.2.2.3.

8.4.4 Modelling loosely restricted production responses

In Section 7.2.2.4 we analysed a type of task for which topical knowledge is hard to predict given the relation between input and response. We classified this type of task as structured communication, which has a certain degree of unpredictability – as Littlewood (2004: p. 322) puts it.

The criteria for correctness for this task require that the response includes, among others, expressions to show satisfaction or dissatisfaction with a product, and to ask information about a product.

For the recognition of this type of linguistic structure, one could use recognition paths as the ones reflected in Figure 8.18.

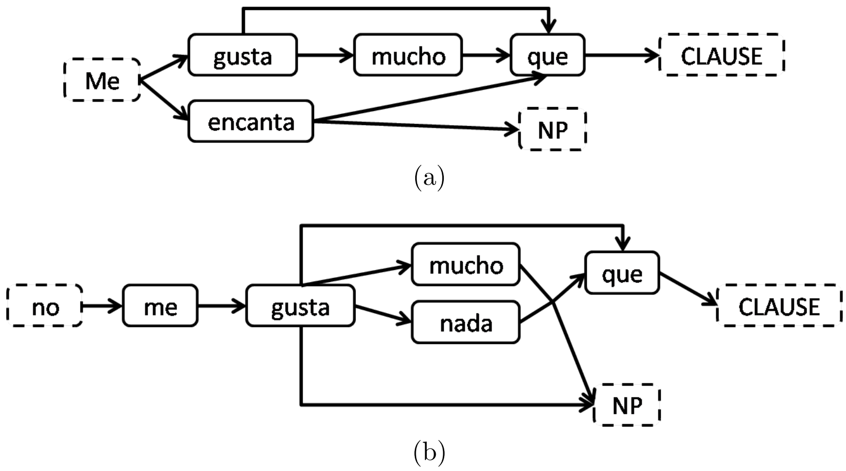


Figure 8.18: Language content recognition paths for loosely restricted responses.

Note however that the ending nodes correspond to linguistic elements identified by its syntactic class: noun phrase (NP), or subordinate clause (CLAUSE). With this type of recognition strategy it would be possible to evaluate the correctness of the response in terms of linguistic knowledge. One could think of other sorts of linguistic resources, particularly resources including semantic categories, such as ontologies, to ensure that (at least part of) the information included in the analysed linguistic elements (NP or CLAUSE) is related with candies, food, or, for instance, teeth health-related topics.

The automaton to be included in the FG module for the evaluation of responses at a global level is reflected in 8.19. This automaton allows for the recognition of response elements that are indicators of asking for further information, or expressing an opinion. However, we cannot be sure that the semantics of such questions or opinions are coherent with the task's goals.

8.5 Automatic generation of summative feedback

The distinctive feature of summative feedback in our pedagogical research setting is that it combines information from different levels of analysis to produce pedagogically informed quantitative assessment.

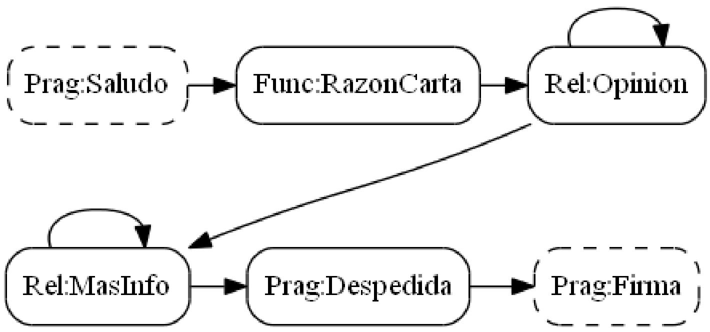


Figure 8.19: Recognition path of the response components of a task with a loosely restricted response – Activity 5 in Subtask 2 of *Antención al cliente* in ALLES.

8.5.1 Analysing learner responses for the generation of summative feedback

As described in Section 8.2.2, summative assessment in our research context focuses on four dimensions: communicative contents, lexical contents, sentence structure and accuracy, overall fluency and quality of the text. In practical terms, these four dimensions are correlated with textual characteristics that can be extracted automatically from the analysis that automatic language processing tools yield. We will review how each dimension is obtained and quantified. However, recall that the SALA and SFGL specifications of the information required is reflected in Tables 8.5, 8.6, 8.7 and 8.8.

Analysing the communicative contents of learner responses Communicative contents are based on counts of two different features of the RIF: the thematic contents, and the pragmatic contents (which are part of the linguistic contents). While the former are related to the task's topic, the latter are related to the text genre of the task's response. The thematic contents counts are based on analysis codes such as the ones described in Table 8.5: *Rel:YourDepartment*, *Rel:DesiredCourse*, *Rel:UsefulFutProjects* and so on. The linguistic contents counts are based on analysis codes such as the ones in the Pragmatics block in Table 8.6 *Prag:Greeting*, *Func:IntrodYourself*, and so on.

Analysing the lexical contents of learner responses The dimension *lexical contents* takes into account the total number of words in the response, and the use of specific vocabulary in comparison with the use of specific vocabulary in the model responses provided by either teachers or native speakers for exactly the same text. This implies the following processing tasks: word tokenisation, word lemmatisation and POS tagging. POS tagging is used, because as we described, only so-called concept words (that is, adjectives, nouns and verbs) are taken into account. Moreover, it requires the identification of the domain-relevant words, that is, the words that according to the pedagogical goals of the task are specific vocabulary. As we described in commenting on Tables 8.6 and 8.7, this requires either an additional lexicon or some more information on the lexical entries of the standard lexicon. The evaluation of the use of specific vocabulary is based on counts and presence of so-called content words (nouns, verbs and adjectives) in the learner response in comparison to their use in reference models. For more details on model building and reference comparison see Appendix D.

Sentence structure and accuracy The assessment of this dimension takes into account the ratio of complex/simple sentences, the number of discourse markers in the response, and the number of grammar and usage errors. To do so the text has to be processed with the tokenisation module, which provides sentence segmentation, and main and subordinate clauses need to be identified exploiting the syntactic information provided by the morphosyntactic module. The strategy adopted for the identification of discourse markers is to mark them during the dictionary look-up process by means of an ad-hoc code mapping process that checks a pre-defined list of discourse markers. As for the detection of grammar and usage errors, these are provided by the grammar checking module. The analyses provided be each of these modules are the basis for the counts that determine the generated feedback.

Overall fluency and quality of the text The dimension *Overall fluency and quality of the text* is based on counts on sentences or paragraphs and the ratio of non-word spelling errors. The number of sentences or paragraphs is provided by the tokenisation module, and the number of non-word spelling errors by the spell checker.

8.5.2 Evaluating learner responses for the generation of summative feedback

To provide learners with the grades and feedback messages reflected in the specifications in Tables 8.3 and 8.4, a relational database is built. The database contains the tables and the data structure necessary to correlate the dimensions of the assessment, and maps linguistic indicators with concrete feedback messages.

Figure 8.20 shows the mapping of the linguistic indicators for a response obtained from a learner in a trialling session. The figure shows an XML file that contains all the necessary numerical indicators and is used as input for the query in the database. The XML syntax is quite simple: The relevant tags are $< exText >$ and $< exResult >$. $< exText >$ has two attributes: *exId* identifies the activity so that the evaluation models can be determined, and *exAnswereId* allows for the identification of the learner, the learner attempt, and so on.

The tag *exResult* contains the values of the linguistic indicators organised in "blocks" and "subblocks". Each of the assessed dimensions corresponds to a block: 1 corresponds to communicative contents, 2 to lexical contents, 3 to syntactic structure and accuracy, and 4 to overall fluency and quality of the text. In turn, each of the dimensions has as many sub-blocks as indicators defined in the specifications: two for all of them, except for the dimension syntactic structure and accuracy, which has three subblocks (see Section 7.2.2.3.2).

As for the particular values in Figure 8.20, the values for block 1, communicative contents, indicate that four out of the six expected response elements in terms of thematic contents were detected, while all four elements expected in terms of linguistic contents were found.

Block 2, lexical contents, presents a value of 98.7% for subblock 1, use of specific vocabulary, and a value of 119 for subblock 2, number or words. The latter figure is the total number of words, in this case 119, a simple measure of fluency that will be compared with a reference value provided by task designers. The other figure, the percentage, is obtained by comparing the presence of domain words in the learner response to their presence in the so-called reference model. For the purposes of this experiment, the reference model for the activity of the email registration, whose feedback specifications are described in Section 8.2.2, is a set of three texts produced by proficient non-native speakers of English. Appendix D describes this comparison procedure in detail.


```
<?xml version="1.0"?>
<allesAssessment>
   <exText exId="4" exAnswerId="6688">
      <exResult block="1" subblock="1" value="4">
      <exResult block="1" subblock="2" value="4">
      <exResult block="2" subblock="1" value="98.70">
      <exResult block="2" subblock="2" value="119">
      <exResult block="3" subblock="1" value="9">
      <exResult block="3" subblock="2" value="8">
      <exResult block="3" subblock="3" value="0">
      <exResult block="4" subblock="1" value="9">
      <exResult block="4" subblock="2" value="0">
   </exText>
</allesAssessment>
```


Figure 8.20: Indicators obtained from the learner response for summative assessment.

Block 3, the dimension sentence structure and accuracy, has three subblocks: Subblock 1 is related to the number of sentences in the response; subblock 2 is related to the number of discourse markers; and subblock 3 is related to the number of grammar and usage errors in the text detected by the system. All values are compared with reference values provided by task designers.

Block 4, overall fluency and quality of the text, presents two subblocks: Subblock 1 corresponds to the number of sentences for this activity. Subblock 2 corresponds to the number of spelling errors resulting in non-words detected in the text. All values are again compared with reference values provided by task designers.

With the XML file in Figure 8.20 and the specifications in Tables in Section 8.2.2, the feedback that the learner would obtain is reflected in Figure 8.21.

8.6 Chapter summary

In this chapter, we introduced the Automatic Assessment Specification Framework, and its two components the Specifications for Automatic Linguistic Analysis, and the

General assessment	Grade
Careful: there is some information missing in the exercise.	2 / 4
Excellent. Your text reads well and is precise. You are using the adequate vocabulary.	4 / 4
Good. Your text is adequate, but there some minor errors.	3 / 4
Your text has a somewhat adequate structure and there are no spelling mistakes.	3 / 4

Content checking

1. Specify whether your course fits with your schedule!
2. Specify which courses you want to take in the future.

Figure 8.21: Feedback generated by the latest version of ALLES for a particular learner response.

Specifications for the Feedback Generation Logic. We showed how domain-specific requirements for the NLP-based analysis and the feedback generation logic can be produced on the basis of RIF-based thematic and linguistic content characterisation of ICALL tasks. The AASF is the framework that provides a pedagogically-informed and NLP-oriented characterisation of ICALL tasks in linguistic and assessment terms, and it plays an important role in the connection between FLTL and NLP.

We showed how the specific requirements of the automatic assessment module are actually implemented in our ICALL research setting. Following a finite-state shallow approach to NLP we can model domain-specific strategies for the generation of formative and summative assessment on the basis of automated linguistic analysis of learner language. To provide with this assessment strategy we presented the Linguistic Analysis module, particularly the Information Extraction Modules, which identify linguistic and communicative elements relevant for the evaluation of the analysed responses. We also presented the Feedback Generation module, which searches for the communicative and linguistic characteristics of learner responses that support the judgements on the correctness/incorrectness of the response.

The implementation of the rule-based approach was exemplified by showing the conversion of pedagogical requirements into specific rules for handling well-formed and correct responses, as well as ill-formed and incorrect responses. The solution illustrates existing NLP tools enhanced with manually written resources can be custom-tailored to the pedagogical needs. This adaptation results in the implementation of form-based strategies for the analysis of meaning in a particular domain.

Finally, we exemplified how the feedback generation logics defined through the SFGL has been implemented in our research setting. We saw how the creation of finite-state rules at a more abstract level of representation allows for a strategy for the evaluation of responses at a global level. This is the basis of the formative assessment strategy. The implementation of the summative feedback generation strategy used similar NLP-based linguistic indicators, but by means of a relational database system allowed for a correlation of measures that produced pedagogically informed summative assessment.

Chapter 9

ICALL task complexity on the basis of learner data

In the previous chapters of Part III, we described the use of a design-driven methodology for the pedagogical characterisation of ICALL tasks to inform the implementation of an NLP-based feedback generation strategy. This chapter introduces learner data as a variable in the development of this feedback generation strategy. Learner data is a critical source of information for a proper understanding of the characteristics of the language elicited from learners.

Our aim is to use learner data to compare actual learner responses with the RIF-based envisaged responses. This will provide linguistic evidence of how similar or how different the thematic and linguistic contents of learner responses are compared to those specified in the RIF analysis. Through this comparison we will learn about the nature and the amount of variation found in actual learner responses with respect to designer expectations.

We will perform this comparison by using RIF terminology with the goal to inform both FLTL and NLP experts. Since the RIF characterises activities in terms of language, this will be a natural meeting point for those who are interested in language as a system to be taught, and those who are interested in language as a system to be computationally processed. As a result of the comparison, we will be able to evaluate the extent to which actual learner responses correlate with pedagogical goals and the extent to which the NLP-based assessment might be able to handle learner responses accordingly.

To achieve this goal, we analyse learner responses to three of the ICALL tasks presented in Chapter 7, and we analyse quantitatively and qualitatively the characteristics of learner responses using the RIF. This analysis presupposes an annotation methodology that we present. After this analysis, we discuss the effects that response length and response variation have on the complexity of ICALL tasks in terms of NLP as a means to judge their computational feasibility. Finally, we exemplify two corpus-driven strategies for the adaptation of the modules for the processing of learner language to the domain and to the language characteristics.

9.1 Annotation of learner responses

Our goal here is to characterise learner variation on the basis of a comparison of actual learner responses compared to RIF-based design specifications of ICALL tasks. In analysing variation we focus on two aspects: The first one is the extent to which learner responses actually match with or diverge from the RIF specifications in thematic and linguistic contents. The other aspect is the formal correctness of learner language, that is, the presence of well-formed or ill-formed structures in learner responses. This section presents the annotation criteria and the annotation scheme.

9.1.1 Comparing design-based specifications and learner responses

The annotation process aims to determine if learner responses include or not language envisaged by designers for the implementation of the assessment module. On the basis of design specifications we model both well-formed correct responses and responses including deviant structures – see Section 8.4, and in general in Chapter 8. Therefore, learner responses containing variation might result in inappropriate responses or ill-formed language, but they might also not.

The annotator's task will be to indicate whether the thematic and linguistic contents of a particular learner response matches with the design-based specifications. If the learner's response matches one of the specified language patterns, then it is marked as a *Match*. If it does not, it is marked as an *Alternative*.

The responses in (15) and (16) exemplify annotations classified as a *Match*, since their linguistic structures match with design-based patterns – included in italics below the actual responses. The response in (15) matches exactly with one of the patterns according to specifications.

(15) What improvements would you like to see in the Stanley Broadband service?
What improvements would you like to see in the Stanley Broadband service?
→ Match

In contrast, the sentences in (16) contain responses that match with one of the design-based patterns, even if they present ill-formed linguistic structures – marked with an asterisk and in curly brackets. The thematic contents of the responses are the expected ones, the linguistic contents are not, but the variation was envisaged through simple transformation operations on the model of the well-formed versions of the responses. While the response in (16a) presents variation that affects only one word, the missing determiner, the one in (16b) presents variation that affects two words: the missing determiner and the pluralisation of the noun *service*.

(16) a. What improvements would you like to see in *{Stanley Broadband service}?
What improvements would you like to see in $\emptyset_{Det}$ *Stanley Broadband service?* → Match

b. What improvements you would like to see in *{Stanley Broadband services}?

What improvements would you like to see in $\emptyset_{Det}$ *Stanley Broadband* LEMMA:*service?* → Match

The response in (17) is an example of annotation classified as an *Alternative* response – the deviation with respect to the specified patterns is highlighted using small capitals. The two specified responses (in italics below) are quite different from the learner's response in terms of lexical and syntactic contents. The response does not include a reference to the *askee* (you is missing), and the *improvements* become the subject of the sentence. Moreover, the service in question, *Stanley Broadband*, is not mentioned, a general reference to customer satisfaction is introduced.

(17) What improvements SHOULD BE INTRODUCED TO ENHANCE CUSTOMER SATISFACTION?
What improvements would you like to see in the Stanley Broadband service? → Alternative
What improvements would make more people want to subscribe to Stanley Broadband? → Alternative

9.1.1.1 Correctness and well-formedness

In addition to the matching between learner responses and design-based specifications, we analyse also the correctness and the well-formedness of learner responses. We distinguish between response fragments or whole responses that are correct under pedagogically motivated criteria for correctness, and response fragments that are well-formed or ill-formed linguistically speaking. The reason to distinguish these two types of annotation levels corresponds with the distinction between focusing on meaning or focusing on form.

Therefore, responses and response fragments will be classified as:

Correct Texts that accomplish the activity's criteria for correctness, that is, that express a concept that is expected in the response according to specifications.

Incorrect Texts that are inappropriate under pedagogical criteria, that is, that express a concept unacceptable as part of the response according to specifications.

Well-formed Texts that are linguistically speaking grammatical.

Ill-formed variation Texts that are linguistically speaking ungrammatical.

Correct Well-formed	Yes	No
Yes	Correct & Well-formed	Incorrect & Well-formed
No	Correct & Ill-formed	Incorrect & Ill-formed

Table 9.1: Classes of variation obtained by crossing the criteria of correctness and well-formedness.

As a result of crossing the two classification criteria, we expect four types of variation as reflected in Table 9.1. Note this classification is orthogonal to the *Match/Alternative* classification: responses that match with specifications can be correct and well-formed as (15), correct and ill-formed as the ones in (16), or incorrect and well-formed, or incorrect and ill-formed. Responses that do not match with specifications can also be included in any of the previous four classes: (17) is an example of a incorrect and well-formed alternative response.

9.1.2 Scheme for the annotation of learner responses

We use the following scheme to annotate the above mentioned information:

1. Each response is annotated using the XML tag `<resp>` with the following attributes:

   (a) `id`: the identification number of the response.

   (b) `exid`: the identification number of the ICALL activity

   (c) `item`: the item number within the activity (at least one)

2. For each response, annotations are marked with the XML tag `<ann>`. Each `<ann>` tag contains the following attributes:

   (a) `match`: this attribute indicates whether that fragment matches with the design-based specifications. Its possible values are *yes* or *no*.

   (b) `corr`: the attribute *correctness* indicates whether the response or response fragment is correct (CO) or incorrect (IN).

   (c) `form`: this attribute indicates whether that fragment is linguistically well-formed (WF) or ill-formed (IF).

   (d) `know`: classification of the variation in terms of linguistic knowledge (and then the values GR, grammatical, TX, textual, FC, functional, and SL, sociolinguistics are used) or as topical knowledge (and then the value TK is used).[1]

   (e) `subknow`: Sub-classification of the variation within the knowledge type. Each knowledge class has different subknowledge classes:

   - GR: graphology (GRA), morphology (MOR), syntax (SYN), or semantics (SEM).
   - TX: cohesion (COH), or rhetorical organisation (RHE).
   - FC: ideational (IDE), manipulative (MAN), heuristic (HEU), imaginative (IMA).
   - SL: dialect or variety (DIA), register (REG), idiomatic expressions and naturalness (IDM), and figures of speech and cultural references (FOS)

[1]Note that Bailey and Meurers (2008)'s *content* and *linguistic form* correspond with Bachman and Palmer (1996)'s *topical knowledge* and *linguistic knowledge*. Strictly speaking the respective definitions are not identical, but notionally they are largely coincident concepts.

- TK: empty (ETY).

(f) `trans`: type of transformation according to the text surface: alternate choice (AC), alternate order (AO), additional element (DE), omitted element (OE), blending structure (BS), other (TH), and several (SV).

(g) `re`: identification of the part of the response in which the variation is located. This corresponds to the response elements in which responses are divided in the Linguistic Analysis and Feedback Generation modules.

The attributes *know, subknow, trans* and *rc* provide fine-grained information about the pedagogical and linguistic nature of the performed annotations. While the attributes *know* and *suknow* are related with the RIF and Bachman and Palmer (1996)'s characterisation of the thematic and linguistic contents of responses, the attribute *trans* is in line with the tradition of classifying errors according to the surface transformation operation on the text (Damerau, 1964; Kukich, 1992). However, we use more neutral terms, so that it can be used to identify both ill-formed and well-formed variation. Finally, the attribute *re*, response element, is a RIF/AASL internal localisation of the annotation with respect to the nodes of the finite-state rules in the Feedback Generation module (see Sections 8.4.1.2 and 8.4.3).

9.2 Learner language in task responses

In this section, we analyse the annotations made on three sets of learner responses to the tasks exemplifying Task Types I, III and IV presented in Chapter 7.

9.2.1 Responses to a Type I task

In this section we analyse the responses from learners of English as a foreign language to the ICALL task corresponding to Activity 4 of Subtask 3 of the learning unit *Customer Satisfaction and International Communication*. The pedagogical characteristics of this tasks were described in Section 7.2.2.1. The implementation of the NLP resources to analyse learner responses automatically were described in Sections 8.4.1 and 8.4.2. Appendix E shows the detailed specifications for the IE modules and FG module.

The responses analysed were obtained from a group of learners at the Heriot-Watt University in Edinburgh in the spring of 2008. The task was done by seven learners as a complement to their face-to-face instruction: Every learner could do the task and other materials in the learning unit on his or her own and voluntarily.

Learner ages are between 20 and 35 years old. Their mother tongues were Arabic, Urdu, Galician/Spanish, Polish, Japanese and German. According to the person who recruited them they were all B2 level learners. According to the profile questionnaires, learners had all learned English for more than five years, and all of them used computers on a daily basis for many tasks, among them studying, working, searching for information, entertaining themselves and shopping.

Activity item	R	SP	Matches		Altern.	
			Patt.	Inst.	Patt.	Inst.
Item 1	7	1	1	2	4	5
Item 2	6	2	2	4	2	2
Item 3	6	2	0	0	6	6
Item 4	5	1	1	5	0	0
Item 5	5	2	1	3	2	2
ALL	29	8	5	14	14	15

Table 9.2: Manually annotated well-formed variations per response component.

9.2.1.1 Qualitative analysis of the language of learner responses

We first look at the matching between envisaged responses and learner responses. After that we comment on the correctness of the responses, independently of the fact that they match with the envisaged specifications. Next we look at the distribution of well-formed and ill-formed variation in learner responses and we present this information segmented by matching of the response with envisaged specifications and by correctness of the response. Finally, we comment on the different levels of linguistic knowledge for which matching and variation phenomena was found.

Matching between envisaged responses and learner responses

The task in question consisted of five different items, and for each of them learners had to provide a limited production response. Table 9.2 shows the number of matches and alternative structures annotated for each of the task's items, one item per row. The first column shows the number of actual *R*esponses; the second one, *S*pecified *P*atterns, shows the number of well-formed patterns specified in the RIF; the third and the fourth column show the number of matches and alternatives found.

As for the *Match* and *Altern*ative structure columns, they are divided into two further columns. The *Patt*ern column and the *Inst*ance column: The former shows the number of single patterns used by different responses, while the latter shows the total number of responses that actually use that pattern.

Item 4 is the only task item in Table 9.2 for which all responses provided by learners use one of the envisaged patterns. In contrast, none of the responses provided for Item 3 uses one of the envisaged patterns. As for Item 1, the responses reflect a notable mismatch with design-based specifications: out of seven responses, two of them follow an envisaged pattern, and five do not use an envisaged pattern. Moreover, among the five responses identified as alternatives, four different patterns are observed. As for the responses for the other two items, they present a slightly more balanced proportion in favour of the responses using envisaged patterns: 4:2 for Item 2, and 3:2 for Item 5.

A qualitative analysis of the different responses for Item 1 shows minor formal variation in the two responses that match with one of the specified patterns, and interesting formal variation in non-envisaged responses.

The responses in (18) and (19) match with one of the specified patterns allowing for ill-formed variation. While in (18) there is a determiner and the final question mark missing, in (19) the word *satisfied* is spelled as **satsfied*, the name of the product, *Stanley Broadband*, is omitted, and the expression **you recieve* is unexpectedly included. Despite both responses match with design-based specifications, only the response in (18) is considered correct, and (19) is considered incorrect because it fails to comply with some of the key criteria for correctness.

(18) How satisfied are you with STANLEY$_{DetOmitLeft}$ Broadband SERVICE$_{QuestionMarkMiss}$
How satisfied are you with $\emptyset_{Det}$ Stanley Broadband service $\emptyset_{?}$ → Match

(19) How *{satsfied} are you with the service you *{recieve}?
How RESPELL:satisfied are you with the $\emptyset_{SB}$ service UNEXPECTED? → Match

As for the responses that do not match with any of the specified patterns, we observe cases as the ones in (20) and (21), where the matching fails even though their wording is very close to some of the envisaged patterns. This is due to the fact that certain word order transformations were not foreseen. For instance, (20) presents the word *satisfied* in the wrong place, not after *how*, but after *are you*. In (21) the word *satisfied* is also in the wrong position, but in addition there is the word *much* after *how*, and an inversion in the order of the subject and the verb, *you are* instead of *are you*. Though both the addition of extra words between *how* and *satisfied* and the lack of inversion of subject and verb are envisaged, the misplacing of the word *satisfied* prevents the matching with the design specifications.

(20) How ARE YOU SATISFIED with Stanley Broadband?
How satisfied are you with Stanley Broadband service? → Alternative

(21) How MUCH you are SATISFIED with Stanley Broadband?
How UNEXPECTED satisfied INVORDER:you_are with Stanley Broadband service? → Alternative

As for the other three responses that do not much with specifications, we present and analyse them in (22) through (24). The three of them present linguistic structures used as alternatives to the envisaged ones. (22) and (23) present similar lexical items and lexical roots (cf. *satisfy* and *satisfaction*), but they introduce verbs such as *is* and *rate* that lead to very different lexico-syntactic patterns. In addition (23) gives the intended customer the opportunity to numerically rate his or her satisfaction, something originally not envisaged.

(22) How IS YOUR LEVEL OF SATISFACTION with Stanley Broadband?
How satisfied are you with Stanley Broadband? → Alternative
How happy are you with Stanley Broadband? → Alternative

(23) How WOULD you RATE YOUR SATISFACTION with Stanley Broadband FROM A SCALE OF 1 TO 5?
How satisfied are you with Stanley Broadband? → Alternative
How happy are you with Stanley Broadband? → Alternative

Item	R	Correct	Incorrect
1	7	3	4
2	6	4	2
3	6	2	4
4	5	5	0
5	5	2	3
ALL	29	16	13

Table 9.3: Distribution of correct and incorrect responses to the items in Activity 4 of Subtask 3 in *Customer Satisfaction and International Communication.*

As for (24), it presents a totally different lexico-syntactic pattern. The use of a different lexical choice to ask about someone's satisfaction, *to feel*, requires a different auxiliary verb, as well as a different preposition.

(24) How DO YOU FEEL ABOUT Stanley Broadband?
How satisfied are you with Stanley Broadband? → Alternative
How happy are you with Stanley Broadband? → Alternative

Note these last three groups of examples show instances of well-formed variation in correct responses that did not match with the design specifications, and that in terms of the linguistic contents subcategories in the RIF variation is observed at the level of functional, syntactic and lexical content.

Correctness of responses

Table 9.3 shows the distribution of correct and incorrect responses for each of the task's items. If we compare the figures in the columns of instances (labelled *Inst.*, columns five and seven) in Table 9.2 with the figures in Table 9.3, we observe that the distribution of correct and incorrect responses is not always the same as the distribution of responses using an envisaged or a non-envisaged pattern. It is the same for Items 2 and 4, but not for Items 1, 3 and 5. This suggests that response correctness is not correlated with the ability to predict learner responses in the design phase, which argues in favour of the design of NLP strategies for the analysis of incorrect responses as much as for the analysis of correct responses. Of course, it argues also in favour of corpus-driven approaches, that is, approaches taking into consideration actual learner responses.

Interestingly too we observe that for Item 3, for which none of responses followed an envisaged pattern according to Table 9.2, the proportion of correct and incorrect responses is 2:4. In our opinion, this suggests that the mismatch between design-based specifications does not always co-relate with poorer performance of the learner.

Well-formed and ill-formed language

In this section we analyse the distribution of well-formed and ill-formed linguistic structures corresponding to variation in the set of responses of this task. Table 9.4 shows the number of well-formed and ill-formed annotations for the different possible

Item	Envisaged				Not envisaged				Total	
	Correct		Incorr.		Correct		Incorr.			
	WF	IF	WF	IF	WF	IF	WF	IF	WF	IF
1	0	2	0	2	3	2	0	6	3	12
2	0	0	1	1	1	0	0	2	3	4
3	0	0	0	0	2	0	0	8	2	8
4	0	7	0	0	0	0	0	0	0	7
5	0	0	0	0	0	0	2	2	2	3
Total	0	9	1	3	6	2	2	18	10	34

Table 9.4: Well-formed and ill-formed structures in correct and incorrect responses to the items in Activity 4 of Subtask 3 in *Customer Satisfaction and International Communication* in ALLES.

combinations of the other two criteria: envisaged correct and incorrect responses, and non-envisaged correct and incorrect responses.

The first two columns reflect the well-formed and ill-formed structures reflecting variation in responses that use envisaged patterns and are correct. The first of them is filled with zeros: No well-formed variation was observed in envisaged correct responses. The second column reflects ill-formed structures in responses that are correct, but whose ill-formed deviations were envisaged. Examples of the responses found here are the ones in (25). In (25a), there is a missing determiner and a word in the wrong capitalisation; in (25b), a determiner and a question mark are omitted. Both responses were considered correct according to the design specifications, even though they do not fully accomplish all the criteria for correctness.

(25) a. How often do you use *{internet}?
How often do you use the Internet? → Match

b. How satisfied are you with STANLEY$_{DetOmitLeft}$ Broadband SERVICE$_{QuestionMarkMiss}$
How satisfied are you with $\emptyset_{Det}$ *Stanley Broadband service* $\emptyset_{?}$ → Match

Note that the envisaged deviations are observed at the level of graphology and syntactic content.

The third and the fourth columns present respectively the number of well-formed and ill-formed annotations in responses that were incorrect, whose patterns were envisaged. The response in (26) exemplifies both well-formed and ill-formed annotations. The use of the incorrect, even if well-formed, expression *the service you receive* is handled by a pattern that envisages two deviations: one that entails the omission of the service name, *Stanley Broadband*, and one that allows for the inclusion of unexpected contents at the end of the response. The spelling error in **satisfied* exemplifies an ill-formed annotation.

(26) How *{satsfied} are you with THE SERVICE YOU RECEIVE?
How RESPELL:satisfied are you with the $\emptyset_{SB}$ service UNEXPECTED? → Match

The variations observed occur at the levels of graphology (the spelling error), and at the level of semantics: Arguably, the reference of the definite descriptions *the* *service you receive* and *the Stanley Broadband service* are the same, so we consider them different meanings to referring to the same entity in that context.

The fifth and the sixth column show the number of annotations in responses that are correct but whose pattern was not envisaged. The response in (27) exemplifies a non-envisaged well-formed correct response.

(27) What is the least satisfying feature of Stanley Broadband?
What do you like least about Stanley Broadband?→ Alternative
What feature of Stanley Broadband do/did you like least?→ Alternative

The one in (28) shows a response that contains a spelling error *de* (instead of *do*). What makes the response not envisaged is not the spelling error in *de*, but the fact that the pattern *do you feel about X* was not envisaged.

(28) How *{de} you feel about Stanley Broadband?
How satisfied are you with Stanley Broadband? → Alternative
How happy are you with Stanley Broadband? → Alternative

Finally, the seventh and eighth columns show the annotations made in incorrect responses that were not envisaged. Well-formed annotations in this kind of response are exemplified in (29). Despite being linguistically well-formed, the response does not include a reference to the product *Stanley Broadband*, and is considered incorrect.

(29) What improvements SHOULD BE INTRODUCED TO ENHANCE CUSTOMER SATISFACTION?
What improvements would you like to see in the Stanley Broadband service? → Alternative
What improvements would make more people want to subscribe to Stanley Broadband? → Alternative

Samples of responses that were incorrect and contained ill-formed variation that was not envisaged were show previously in (20) and (21). In those sentences, unexpected variation originates in the placement of the word *statisfied*.

Linguistic knowledge types and variation

In this last section, we draw our attention to the quantitative and qualitative presence of variation from different types of linguistic knowledge according to the RIF. Out of the 41 annotations marked for the set of responses to this activity, nine of them correspond to well-formed variation annotations and 32 of them to ill-formed variation annotations.

Regarding the annotations of well-formed variation, all of them are under the class grammatical knowledge. The response we saw in (23) is an example of response containing instances of well-formed variation at the level of grammatical knowledge, since the concept is expected in the specifications, but not in that form.

Out of the 34 ill-formed variation annotations, 29 of them correspond to annotations at the level of grammatical knowledge and five to annotations at the level of sociolinguistic knowledge. The former are divided into 13 annotations at the syntactic level, ten at the graphology level, three at the semantic level, and another three at the syntactico-semantic level.

The sentences in (20) and (25) were examples of responses containing ill-formed variation at the syntactic level. (20) exemplifies a wrong ordering of the elements in the sentence; (25) exemplifies an error where a determiner is omitted in front of the word *Internet*. Another example of ill-formed variation at the syntactic level is the response in (30), a sentence where the auxiliary *do* has been omitted.

(30) What improvements *{you think} we should make to our service?

Examples of ill-formed variation at the level of graphology were presented before in (25), (28), and (19). (25) presents a capitalisation error that results into a word: *Internet* vs. *internet*, while (28) and (19) present both spelling errors resulting in non-words.

As for ill-formed annotations classified as errors at the semantic level, we find the ones in (31). All the sentences in (31) present problems in the wording of the expressions containing the superlative. According to annotations, there is a problem that is somewhere between the semantics and the naturalness of the expression – compare each of the sentences with the hypothesised target responses, below in italics, the target response for (31a) is under (31b).

(31) a. What is *the least good feature* of the Stanley Broadband service?
b. What is *your least favourite feature* of the Stanley Broadband service?
What is the least satisfying feature of the Stanley Broadband service?
c. What is *your best favourite feature* of the Stanley Broadband service?
What is the best feature of the Stanley Broadband service?

As for the annotated ill-formed variation at the level of sociolinguistic knowledge, we observe responses as the one in (22), which is marked as unnatural. A second example is the response in (32), where the use of the word *thing* is considered not precise enough, too informal for the task setting.

(32) What is the $^{\#}${thing} you like the least?

Analysis summary

To sum up, the analysis of the responses for this task shows that five of the envisaged patterns accounts for 14 of the analysed responses, whereas 14 different patterns were identified in the 15 responses that did not follow any of the envisaged patterns. We interpret this as an indicator that design-based specifications can be good and effective. However, this evidence suggests that corpora-driven approaches are necessary for the development of assessment strategies for ICALL tasks.

As for the correctness of the analysed responses, 16 of them were correct and 13 were incorrect, and, as we said, the correlation between correctness and matching with the envisaged responses is not high. Therefore, we think that this justifies the design of NLP strategies for the analysis of incorrect responses as much as for the analysis of correct responses. This evidenced further supports the need of corpus-driven approaches.

Finally, in the analysis of the types of linguistic knowledge involved in the observed variations, we saw that most of them were classified as grammatical knowledge, while only a few of them were classified as sociolinguistics knowledge. We interpret this and the absence of annotations classified as a divergence at the level of functional or textual knowledge, as a sign that the type of task really presents a direct and narrow relationship between input and response. At the same time, from a pedagogical perspective, this task seems to foster the use of specific constructions, thus, it seems coherent to classify it as communicative language practice in Littlewood (2004)'s terms (see Section 7.2.2.1).

9.2.2 Responses to a Type III activity

In this section we analyse the responses from learners of English as a foreign language to the ICALL task corresponding to Activities 1 and 2 of the Final Task in *Education and Training*, described in Section 7.2.2.3 as a sample of Task Type III. This activity requires learners to write an email to register for a course on the basis of specific input data.

The responses analysed were obtained at the Universidad Europea de Madrid and at the Universitat Pompeu Fabra in Barcelona in the spring of 2005. There was a total of 14 participants and the materials were used as a complement to their face-to-face instruction: Every learner could do the unit on his or her own and voluntarily.

Learner ages were between 20 and 28 years old and their mother tongues were Catalan and/or Spanish. Learners were required to do the DIALANG test and they showed to be either B1 or beginner B2 level learners.[2] According to an initial learner questionnaire, all participants had been learning English for more than five years; in general, they used computers on a daily basis for many tasks, among them studying, working, and searching for information.

9.2.2.1 Qualitative analysis of the language of learner responses

We describe the characteristics of the set of responses to this task in terms of the matching between envisaged responses and learner responses, correctness, distribution of well-formed and ill-formed variation, and the different types of linguistic knowledge observed in the annotations. For this task, we present the data segmented by response elements, not item, due to the fact that this task is responded in one full text, not in items. Analysing the responses at the level of response element facilitates the observation of variation at a linguistic level that corresponds with the structure of the finite-state rules in the FG module. See Section 7.2.2.3 for the pedagogical specifications of this task, and see Section 8.2.2 for its NLP and automated feedback specifications. Appendix E shows the detailed specifications for the IE modules and FG module.

[2] DIALANG is language diagnosis system developed in accordance with the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). URL: http://www.lancs.ac.uk/researchenterprise/dialang/about.

Activity item	R	SP	Matches		Altern.	
			Patt.	Inst.	Patt.	Inst.
Greeting	14	3	1	3	4	11
IntroYourself	13	1	1	6	1	7
YourDept	13	2	1	3	7	10
Course	14	2	1	2	7	12
Schedule	11	2	0	0	9	11
AuthorisedBy	12	2	1	3	6	9
UsefulFuture	12	3	0	0	9	12
FutureInterest	9	2	1	5	4	4
ComplClose	17	1	1	3	5	14
Signature	14	1	1	13	1	1
Unexpected	9	–	0	0	8	9
ALL	138	16	8	38	61	100

Table 9.5: Responses using linguistic structures that match or diverge from design specifications for Activities 1 and 2 of the Final Task in *Education and Training*.

Matching between envisaged responses and learner responses

Table 9.5 shows the number of matches and alternative structures annotated for each of the response elements of the expected answer to this task. The first column shows the number of response fragments identified as belonging to one of the expected response elements. Some response elements might be present more than once in a particular response, which explains why *ComplClose* has 17 counts.

We observe that for most response elements between two and three patterns were specified, see column three, but for all of them only one of the specified patterns was observed at most. Another interesting observation is that for those response elements that belong to the more formal part of the response's the text genre, the email, the ratio of observed patterns and responses tends to be much lower than for the other response elements. And most interestingly this is true both for responses that match with envisaged patterns and for those that do not match, that is, for those using alternative structures, where the ratios are 1:7 for the response element *IntroYourself*, 4:11 for *Greeting*, and 5:14 for *ComplClose*. As for the response element *Greeting*, though the ratio is 4:11 in the responses using non-envisaged patterns, the fact is that out of those 11 responses six of them use the same pattern.

If we observe the ratios of the other response elements (*YourDept*, *Course*, *Schedule*, *AuthorisedBy*, *UsefulFuture*, and *FutureInterest*), we see that *FutureInterest* is the one for which an envisaged pattern was observed most frequently: five out of nine observations. For the rest of response elements the figures are three out of 13, *YourDept*, two out of 14, *Course*, zero out of 11, *Schedule*, three out of 12, *AuthorisedBy*, and zero out of 12, *UsefulFuture*. This reflects a low power of prediction on the side of the specifications for this task and this group of learners.

If we look at the patterns observed in the responses using non-envisaged patterns, the pattern-instance ratio is also quite discouraging: 7:10, *YourDept*, 6:12, *Course*, 9:11, *Schedule*, 6:9, *AuthorisedBy*, and 9:12 *UsefulFuture*, and 4:4, *FutureInterest*.

Item	R	Correct	Incorrect
Greeting	14	2	12
IntroYourself	13	13	0
YourDept	13	11	2
Course	14	13	1
Schedule	11	10	1
AuthorisedBy	12	11	1
UsefulFuture	12	12	0
FutureInterest	9	6	3
ComplClose	17	7	10
Signature	14	14	0
Unexpected	9	6	3
ALL	138	105	33

Table 9.6: Distribution of correct and incorrect responses to Activities 1 and 2 of the Final Task in *Education and Training.*

Correctness of responses

Table 9.6 shows the distribution of correct and incorrect instances of responses elements for the task. The table shows that learners tended to provide correct versions of the corresponding response elements except for *Greeting* and *CompleClose*. For these two, most of the annotations indicate a lack of formality in the learner response as the main reason to consider them inappropriate (more on this below).

The figures in Table 9.6 show a high number of correct responses, while the ones in Table 9.5 show a high number of responses using non-envisaged patterns. This is particularly true for response elements referring to thematic contents, and less for response elements formally restricted by the text genre used in the communicative setting. These two facts support the idea that corpus-driven approaches to ICALL development are the most appropriate strategy for a reasonable handling of variation in learner responses for this activity type.

Well-formed and ill-formed language

In this section we analyse the presence of well-formed variation and ill-formed variation in the set of responses to this task. The analysis of well-formed variation is exemplified for one of the response elements, *Course*, while the analysis of ill-formed variation is presented across all response elements. This distinction in the presentation is due to the fact that well-formed variation for this set of responses was not annotated at levels of description below the response element; therefore, the number of non-envisaged responses is the same as the number of well-formed patterns, after the local linguistic content errors are corrected. Thus, well-formed variation will only be analysed qualitatively.

Well-formed language We focus on the characteristics of the responses including the element *Course*. As shown in Table 9.5, 12 out of the 14 responses including the response element *Course* present linguistic structures alternative to the ones specified, while in Table 9.7 we observed that 11 out of these 12 responses included a correct version of the response element *Course*. In addition, checking back Table 9.5, we see that the 11 responses can be grouped into five different patterns. We analyse how these five patterns differ from the envisaged ones.

To start with, the sentences in (33) show the two patterns envisaged to express the *Course* response element. The main differences are on the way the sentence starts, either indicating a wish with *would like to/ want*, (33a), to or indicating a fact with *have signed up*, (33b).

(33) a. I would like/want to sign up to take/do the course on/called X.
b. I have signed up take/do the course on/called X.

The patterns of the responses using alternative structures to these two patterns are shown in (34), (35), and (36). Responses based on patterns in (34a) and (34b) imply a slight change in the lexical or syntactic choices used to express the wish, and they are used twice and once respectively in the set of the responses.

(34) a. I am planning to take the X course.
b. I want to do the X course.

The patterns in (35) are characterised by expressing the will of participating in the course through the verb phrase *to be interested in*. While (35a) presents a syntax closer to the envisaged patterns (cf. 33), (35b) and (35c) use a more complex linguistic structure including the juxtaposition of two simple sentences. The patterns in (35) are observed four times, (35a), twice, (35b), and once, (35a).

(35) a. I am interested in the course on X.
b. I am interested in (signing up for) one of your courses: namely the one on X.
c. I am interested in one of your courses. I am interested in the course on X.

Finally, (36) is a pattern that includes the registration petition as the reason for the email, which could be inferred from the activity instructions and is coherent with the text type. This pattern, however, is used only once.

(36) I am writing to you to register for the course on X.

Ill-formed language Table 9.7 shows the distribution of ill-formed variation annotation for the set of responses to this task. Table 9.7 shows two interesting results. On the one hand, it shows that ill-formed variation is lower in envisaged responses, independently of their correctness: in total 16 annotations are found in the 38 response elements marked as envisaged, while 76 are found in the 100 response elements marked as non-envisaged (cf. Table 9.5).

The difference in the ratios are approximately one ill-formed annotation every four responses using an envisaged pattern, and three ill-formed annotations every

Resp. Elem.	Envisaged		Not envisaged		Total
	Correct	Incorr.	Correct	Incorr.	
	Ill-formed variation annotations				
Greeting	0	1	1	2	4
IntroYourself	3	0	4	0	7
YourDept	2	0	6	1	9
Course	1	0	9	0	8
Schedule	0	0	8	3	11
AuthorisedBy	1	0	6	2	9
UsefulFuture	0	0	13	0	13
FutureInterest	2	0	2	0	4
ComplClose	1	0	2	9	12
Signature	0	4	0	0	0
Unexpected	0	0	1	5	6
Not in RC	0	1	1	1	3
Total	10	6	53	23	92

Table 9.7: Ill-formed structures in correct and incorrect responses to the items in Activity 4 of Subtask 3 in *Customer Satisfaction and International Communication.*

four responses using a non-envisaged pattern. On the other hand, Table 9.7 shows that ill-formed variation is also present in responses using non-envisaged patterns, independent of the fact that they are correct or not.

Both findings suggest that learners tend to make more errors when manoeuvring with language that was not foreseen by task designers, and might be also correlated with the language learners are exposed to the in the tasks previous to this one. From an NLP perspective, this supports the need for corpus-driven approaches to ICALL system design, as well as the need to handle variation at the level of thematic and linguistic contents.

Linguistic knowledge types and variation

In this section we analyse the types of linguistic knowledge of the ill-formed variation annotations. Out of the 92 ill-formed variation annotations identified 68 are related to grammatical contents, 15 to sociolinguistic contents, and 11 to textual contents.

Within the ill-formed variations related to grammatical contents, there are 22 errors related to syntactic issues, exemplified in (37), (38) and (39).

(37) exemplifies errors of wrong preposition choice: The correct preposition would be *on* in (37a) and *in* or *for* in (37b).

(37) a. I am very interested in your courses: the one **of* Business Communication. AC Wrong Prep
b. I am NAME and I work **at* the Marketing Department.

(38) contains response fragments with errors at the level of the noun phrase. (38a) is a case of wrong choice in the determiner and the noun that results in a noun

phrase with wrong agreement features. (38b) shows a sentence in which a determiner has been omitted before the expression *Marketing Department*.

(38) a. I have no problem to take *this courses*.
b. I am NAME, from **Marketing Department*.

(39) shows an error that has multiple interpretations, and is classified as belonging to two levels of linguistic knowledge: syntax and semantics. Either words are combined incorrectly (or unnaturally), e.g. *projects in the marketing department* would be more natural; or one of the nouns modifying *projects* is unnecessary, e.g. *marketing projects* and *department projects* would both be correct. A third interpretation would be that words are incorrectly chosen, since *marketing department projects* is not the same as *department marketing projects*.

(39) This course will be useful for our #*department marketing projects* by the end of the year.

Another error with this double classification in the semantic and syntactic level is the one in (40). Here the expression *Human Resources* is used with a collective reading, a reading not possible in that context, as opposed to the use of this same expression in a sentence like *Human Resources approves the manager decision*.

(40) Dear #*Human Resources*: (...)

21 ill-formed variation annotations are classified as errors at the level of graphology, most of them typos and wrong use of case, as in (41): **inteltrans*, a proper name not capitalised in (41a); a missing *m* in **Comunication* in (41b) , and a missing *k* in **Than you* in (41c).

(41) a. I am an employee in the Marketing Department of *inteltrans*.
b. I would like to apply for the course on Business **Comunication*.
c. **Than* you very much.

10 of the ill-formed variation annotations are classified as semantic errors, and are exemplified in (42). In (42a) *due to* is expressing a cause, while the hypothetically desired effect is to express an end (*for*) or a relation (*with respect to*). (42b) shows the use of the verb *do* in place of verbs like *sign up* or *take*.

(42) a. This course could be very interesting for my career #*due to* the marketing projects.
b. I would #*do* the course on e-commerce.

9 of the ill-formed variation annotations are classified as morphological errors. (43) exemplifies the wrong choice of the verb modus. In (43a) the form *to*+VERB is used instead of the *ing*-form; (43b) presents an infinitive form instead of the past participle.

(43) a. Would you mind **to tell* me (...)
b. has **encourage* me to go on.

(44) is a wrong choice error regarding the tense of the verb: The past of *will*, *would*, would be better.

(44) It could be an interesting course , and it $^{\#}$*will* help me.

As for ill-formed variation at the level of sociolinguistics, all annotations qualify as wrong uses of the register. (45) shows two of the most common errors in this respect. *Hello* and *Bye* are considered too informal for the setting as expressions to be included in the greeting and the complimentary close.

(45) a. *Hello*, (...)
b. *Bye*, (...)

As for ill-formed variation at the level of textual knowledge, they are exemplified in (46). (46a) is an example of an incomplete expression, the starting of the e-mail, where *Dear* is used without any complement. Even if this could be classified as an error at the level of syntax, a noun missing after an adjective, it was decided to mark it as an annotation at the level of textual knowledge given the specificity of the structure to the text genre. (46b) is an example where a period, instead of the *and*, is required to separate the two clauses.

(46) a. Dear, (...)
b. I see my schedule and the timetable is fine with me *and* I have the autorization of the department manager.

Analysis summary

To sum up, the analysis of the responses for this task shows that eight of the envisaged patterns account for 38 of the analysed response elements, whereas 61 different patterns were identified in the 100 response elements that did not follow any of the envisaged patterns. This finding adds up to what we found for the analysis of the responses to the Type I task in the previous section, and supports the idea that design-based specifications can be good and effective. It also supports that corpora-driven approaches are necessary for the development of ICALL assessment strategies.

As for the correctness of the analysed response elements, 105 of them were correct and 33 were incorrect. Again, the correlation between correctness and matching with the envisaged responses is not high, which supports once more the design of NLP strategies for the analysis of responses with ill-formed variation as much as for the analysis of responses with well-formed variation. This is also an indicator that corpus-driven approaches are the strategy to follow.

Finally, in the analysis of the types of linguistic knowledge involved in the observed variations, most of them were classified at the level of grammatical knowledge. However, in comparison to the responses to the Type I task there is a major presence of deviations classified at the levels sociolinguistics and textual knowledge. Moreover, in the qualitative analysis of well-formed variation there was considerable presence of variation at the level of functional knowledge. We interpret this as an indicator that this type of task presents a the relationship between input and response that is indirect and narrow. This would explain the variation in linguistic structures used by learners to respond to the expected thematic contents, as well as the low observation of variation in the actual thematic contents.

From a pedagogical perspective, this task seems to foster the use of specific constructions, but also the use of specific exponents for language functions with unpredicted linguistic forms. We think this supports the qualification of this task as structured communication practice in Littlewood (2004)'s terms (see Section 7.2.2.3).

9.2.3 Responses to a Type IV task

In this section we analyse the responses from learners of English as a foreign language to the task corresponding to Activity 5 of Subtask 2 in *Atención al cliente*. The activity requires learners to write a letter to a company giving their opinion on a particular product, a type of candy.

The responses analysed were obtained at the Universitat Pompeu Fabra in Barcelona. There was a total of nine participants that used it as complimentary action to their face-to-face instruction: Every learner started the activity in class and could finish it at home, as well as other materials in the course, on his or her own and voluntarily.

Learner ages were between 20 and 28 years old. Their mother tongues were French, German and English. Learners were required to do a DIALANG test and they showed to be beginner B2 level learners. According to learner responses, they all had learned Spanish for at least three years and they used computers on a daily basis for many tasks, among them studying, working, searching for information, and entertaining.

9.2.3.1 Qualitative analysis of the language of learner responses

We describe the characteristics of the set of responses to this task in terms of the matching between envisaged responses and learner responses, correctness, distribution of well-formed and ill-formed variation, and the different types of linguistic knowledge observed in the annotations. Again, we present the data segmented by response elements, not item, due to the fact that this task is responded in one full text; this facilitates the observation of variation at a linguistic level that corresponds with the structure of the finite-state rules in the FG module.

Since the expected response for this task is an extended production response whose relationship between input and response is broad in scope and indirect, in a separate section we will comment on the topics chosen by learners in terms thematic contents. See Section 7.2.2.4 for the pedagogical specifications of this task, and see Section 8.4.4 for the approach proposed to implement and NLP-based feedback generation strategy. Appendix E shows the detailed specifications for the IE modules and FG module.

Resp. Elem.	R	SP	Matches		Altern.	
			Patt.	Inst.	Patt.	Inst.
Saludo	6	2	1	1	3	5
RazonCarta	6	1	1	1	5	5
Opinion	32	9	3	12	13	20
MasInfo	8	3	3	7	1	1
Despedida	4	2	1	2	2	2
Firma	6	1	1	5	1	1
ALL	62	16	10	28	25	34

Table 9.8: Responses using linguistic structures that match and diverge from design-based specifications for Activity 5 of Subtask 2 in *Atención al cliente* in ALLES.

Matching between envisaged responses and learner responses

Table 9.8 shows the number of response elements for which matching or alternative structures are identified in the set of responses. As in the previous analyses, the total occurrences of the response elements in the set of responses is shown in column two, and the number of specified patterns for reach of the response elements in the RIF in column three. We observe that for this task four of the response elements were not observed in several responses: The greeting (*Saludo*), the reason for the letter (*RazonCarta*), and the signature (*Firma*) were missing in three of the nine responses, while the complimentary close (*Despedida*) was missing in five responses.

As for the other two response elements, the one referring to a request of further information, *MasInfo*, is missing in two of the responses, despite its ten occurrences, while the element in which product opinions are expected, *Opinion*, occurs 32 times and is present in all responses. This suggests a tendency in learners to disregard more formal aspects of the activity, since those elements that are more dependent on the text genre were more often left out of the response.

As for the response elements identified as using envisaged patterns, the total number of identified response elements, the instances, is systematically lower than the number of response elements for which a non-envisaged pattern was identified. We relate this finding with the fact that this task was conceived as an open task, one in which structured communication was the goal.

Nonetheless, we also observe that the response elements *Firma* and *MasInfo* do not follow this pattern. For this we find a plausible explanation: On the one side, the response element *Firma* is a very simple linguistic structure, the author's name, so it is simple to predict its structure. On the other side, the good predictability of the response element *MasInfo* might be influenced by the fact that the task's instructions included several samples of questions requiring information as input data, which learners tend to follow.

Appropriateness of responses

Table 9.9 shows the distribution of the observed correct and incorrect response elements. The number of response elements appropriately expressed is very high, 59 out

Resp. Elem.	R	Appropriate	Inappropriate
Saludo	2	6	0
RazonCarta	1	5	1
Opinion	32	31	1
MasInfo	8	8	0
Despedida	4	4	0
Firma	6	5	1
ALL	62	59	3

Table 9.9: Distribution of correct and incorrect responses to Activity 5 of Subtask 2 in *Atención al cliente* in ALLES.

of 62 identified, proportionally the highest for the three activities analysed. Thus, for this task, learners in the given learning context used linguistic expressions that serve the communicative purposes of the text, even if some of the sentences they use, as we describe below, contain some formal errors.

Well-formed and ill-formed language

In this section we analyse the presence of well-formed variation and ill-formed variation in the set of responses to this task. As we will see, the analysis of well-formed variation is exemplified for one of the response elements, *Opinion*, while the analysis of ill-formed variation is presented across all response elements. This distinction in the presentation is again due to the fact that well-formed variation for this set of responses was not annotated at levels of description below the response element. Thus, the number of non-envisaged responses is the same as the number of well-formed patterns, after the local linguistic content errors were corrected. Well-formed variation will be analysed qualitatively.

Well-formed language Table 9.8 shows the analysis of the response element *Opinion*: 12 of the responses used an envisaged pattern, while 20 did not use and envisaged pattern.

Among the 11 responses using an envisaged pattern, six of them use an exponent for the function based on *me gusta* [I like], five of them use one based on *me encanta* [I love], and one of them uses one based on *me parece que* [it seems to me that]. In the observed responses, the first two, *gustar* and *encantar*, are used both to express positive and negative opinions, the latter is only used to express a negative opinion. These are three patterns used from the nine envisaged according to design specifications. The envisaged patterns convey thoughts, believes or points of view by means of expressions such as *creo que* [I believe that], *a mi modo de ver* [the way I see it], *estoy convencido de que* [I am convinced that]. None of the latter was used.

Among the responses using non-envisaged patterns, six of them use an expression based on *X es* ADJ [X is ADJ], where the adjective is sometimes a positive or a negative aspect of the product reviewed. Two other expressions are used by more than once in the set of responses: one based on *encuentro* [I find] as an opinion verb,

Resp. Elem.	Correct	Incorrect	Total
	Ill-formed variation		
Saludo	4	0	4
RazonCarta	5	2	7
Opinion	24	1	25
MasInfo	15	0	15
Despedida	3	0	3
Firma	1	1	2
ALL	52	4	56

Table 9.10: Ill-formed variation annotations in correct and incorrect responses to the items in Activity 5 of Subtask 2 in *Atención al cliente* in ALLES.

and on *(te) permite* [it allows for/you to X]. The other expressions are all used only once: *tiene un sabor X* [its favour is X], *lo mejor es que X* (the best of it is that X), *pienso que es/son para X* [I think that X is/are for X], *te procura* [it gives you X], *hace falta mucho X para Y* [you need a lot of X for Y] as a negative opinion, and so on. Note that all these expressions, except for *pienso que*, are descriptive expressions, expressions that describe the product, but whatever is interpreted as inherently positive or negative in that context.

In our opinion, the data suggest the importance of corpus-driven approaches. The use of statements and descriptions to express opinions is well-known. However, they are not always included as relevant exponents for that function when designing learning materials. One of the reasons for this may be that the interpretation of some of these “hidden” opinions depends more on world knowledge than on communicative or linguistic knowledge.

Ill-formed language Table 9.10 shows the distribution of ill-formed variation annotations in response elements included in correct and incorrect responses. There is a total of 56 ill-formed variation annotations for the response fragments identified. The figures suggest that ill-formed language occurs in learner responses, which again supports the development of robust NLP tools for the assessment of learner responses.

Linguistic knowledge types and variation

In this section we comment on the types of linguistic knowledge classes to which the different ill-formed annotations were classified. Out of the 56 ill-formed variation annotations in the set of responses 41 are classified as errors at the level of grammatical knowledge, nine as errors at the textual level, and six at the sociolinguistics level. As for errors at the level of grammatical knowledge, there are 13 errors each at the level of graphology and at the level of syntax, six at the level of semantics, four at the level of morphology, three of them are classified at the level of both semantics and syntax, and one at the level of both morphology and syntax. We exemplify the three most frequent types of errors at the grammatical level, and a few selected errors at the textual level.

The sentences in (47) exemplify errors at the graphology level. (47a) is an error due to a confusion between *a/ha*, a preposition and the present third person singular form of the auxiliary *haber* [to have]. (47b) is an example of a missing tilde, in this case resulting into a verb form of the verb *practicar* [to practise]. (47c) is an error probably caused by pronunciation, where *s* replaces what it should be a *z*.

(47) a. Quería saber si Smint va **ha* estar comercializado en otros países como en Francia.
I would like to know whether Smint will be commercialised in other countris, such as France.

b. La caja es muy **practica*.
The box is very convenient.

c. Me han salvado la vida más de una **ves* al momento repartir besos.
They saved my life more than once when it comes to kissing.

The sentences in (48) exemplify errors at the syntactic level. There are sentences with agreement errors at the noun phrase and subject-predicate levels, as in (48a), where it should say *los Chupa Chups*, and (48b), where it should say *me encantaron*. There are errors related to missing prepositions, as in (48c), where *a* is missing after *concierne*. And there are errors reflecting wrong use of verb choices: In (48d) the verb phrase should be *os estoy muy agradecido* [I am grateful to you], or else make the sentence something like *cosa que os agradezco mucho* [for which I thank you].

(48) a. Después de **la* Chupa Chups , es otra invención muy ingeniosa.[3]
After Chupa Chups this is again a very witty invention.

b. Me **encantó* todos los diferentes sabores de la gama de caramelo.
I was delighted by the great variety in flavours in this candy product line.

c. En lo que **concierne* los Smint me han gustado también (...)
With respect to Smint, I was delighted with them.

d. Es un placer degustarlos, por lo que **os agradezco mucho*.
It was a pleasure to taste them, for which I am very thankful to you.

The sentences in (49) exemplify errors at the semantic level. In (49a) there is a confusion in the lexical item *como* [eat] that is probably used instead of a word like *trago* [swallow]. In (49b) the expression used *antes que la gente* [before the people] should be *antes que otra gente* [before other people] or *antes que el resto de la gente* [before the rest of the people] because the former includes the speaker and semantically does not make her, the learner, salient as a loyal consumer. Finally, in (49c) there is an example of an invented word. Strictly speaking this is a vocabulary problem: The learner uses a non-existent word, that, in turn, seems to be the result of a morphological blend between *envase* [package] and *embalaje* [packaging]. The errors is classified as a semantic error because it is a lexical choice error.

(49) a. Porque a veces me los $^{\#}$*como* sin querer.
Because sometimes I swallow them by mistake.

[3]This sentence could be corrected in a least two different ways, either replacing *la* with *los*, or adding a preposition *de* to create *Después de la de Chupa Chups...*

b. A lo mejor merezco saberlo #*antes que la gente* por ser tan fiel consumidora.
   Maybe I deserve knowing it before other people do for being such a loyal customer.

c. Encuentro el **envaje* muy conveniente.
   I think the package is very convenient.

As for errors at the textual level they all belong to the subclass coherence errors. The sentences in (50) exemplify them. There are punctuation errors, as the one in (50a), which responds to a convention of letters as a text genre, in Spanish: The greeting usually ends with a colon. There are missing punctuation errors that imply the merging of clauses that end up producing too long, unreadable sentences, as in (50b). A juxtaposition punctuation sign or a coordination conjunction would be needed to repair it. And there is anaphora resolution problems, as the one in (50c), where the definite article *la* [the] is being used in the place of *esta* [this]. In the context, not using *esta* leads the reader to think that the writer is introducing a new entity *gama*, different from the one already mentioned in a previous sentence.

(50) a. Muy señores míos*,*
   Dear Sir or Madam,

b. El nuevo Smint no me gusta nada, su sabor es muy *ácido hace* falta tomarse cuatro a la vez o te lo acabas en un segundo.
   I don't like the new Smint at all, its flavour is too acid. You need to take four of them or they melt in your mouth before you know.

c. Me encantaron todos los diferentes sabores de *la* gama de caramelo.
   I was delighted by the great variety in flavours in this candy product line.

Variation of thematic contents

As we have claimed in the two previous chapters, the development of NLP-based assessment strategies for tasks with less restricted responses as the task representing Type IV tasks is less feasible computationally speaking. Though the linguistic contents of the responses are reasonably predictable, its thematic contents are not. In this section, we analyse qualitatively the variation in the thematic contents of the responses to the Type IV activity analysed in Chapter 7.

Table 9.11 shows the different topics found for each of the three response elements that are more open in terms of thematic contents: *Opinion*, *MasInfo*, and *RazonCarta*. In the *Opinion* elements of the responses we observe learners expressing that they like the candy (7), others praising the packaging (6), the flavours (3), how refreshing it is (2), and then 14 other topics that are mentioned only once. Among these other 14 topics, we exemplify the topics of the size being too small and that of the candy providing you a fresh and clean breath smell.

The other two response elements in Table 9.11, *MasInfo* and *RazonCarta*, present also a variety of topics. Among the topics about which learners ask for further information we observe making the candy bigger, commercialisation in other countries, new flavours, and other three topics occurring once. As for the topics in the element where the reason for writing the letter is stated, they all mention that they have

Response element	Resp. elem. mentions	Differing thematic contents
Opinion	34	LikeThem (7) PackageConven (6) ManyFlavours (3) Refresh (2) TooSmall (1) GoodForDating (1) *Occurring once* (12)
MasInfo	9	BiggerSizeInFuture (2) WillYouCommercialiseAbroad (2) NewFlavourse (2) *Occurring once* (3)
RazonCarta	5	DpProbarEnvioOpin (4) WriteYouAboutSmint (1)

Table 9.11: Well-formed variations for the responses to the items in Activity 5 of Subtask 2 in *Atención al cliente* in ALLES.

tried the candy (which they are asked to do in the instructions) and one of the them says *I am writing you to tell you about Smint*.

In summary, the element for which the number of instances is higher, *Opinion*, shows a couple of topics that seem more frequent: the fact of liking it or not and its packaging. The six topics in the request of further information are fairly equally distributed, and the topics for the element where the reason of the letter is stated seem to be less varied. This analysis suggests that there is variety within topics chosen by learners in the different response elements for which the thematic contents are more open, although some topics tend to be more salient.

Analysis summary

To sum up, the analysis of the responses for this task reflects that ten of the envisaged patterns account for 28 of the analysed responses elements, whereas 25 different patterns were identified in the 34 responses elements that did not follow any of the envisaged patterns. Around 40% of the response elements could be envisaged through design-based specifications, while 60% could not. Moreover, the ratio between patterns and response element instances is 1:3 for the response elements using envisaged patterns, and 5:7 for those using non-envisaged patterns. The figures support the effectiveness of design-based specifications, but also the need for corpus-driven approaches to handle a range of empirically observed linguistic patterns.

As for the correctness of the analysed responses elements, 59 of them were correct and three were incorrect, while in terms of ill-formed variation the total number of annotations observed in those 59 correct response elements is 52, and in the three incorrect ones is four. As it happened with the other two tasks, the correlation between correctness and matching with the envisaged responses is not high, and the presence of ill-formed variation is not correlated with correctness/incorrectness either. This supports once more the design of NLP strategies for the analysis of responses including well-formed and ill-formed variation, as well as the need for corpus-driven approaches.

Still from the perspective of correctness, we saw that for this task there is a higher percentage of correct response elements compared to the other two tasks: for the Type I task it was 16 out of 29, and for the Type III task it was 105 out of 138, compared to the 59 out of 62 in this task. Moreover, we observed a trend in learners to neglect the more formal response elements – tightly related with the text genre.

Finally, in the analysis of the types of linguistic knowledge involved in the observed variations, we saw that most of them were classified as grammatical knowledge, and a few of them were classified as errors at the level of textual knowledge, or the level of sociolinguistics knowledge. In this respect, the levels of linguistic knowledge to which variation annotations are assigned for the responses to this task are similar to those of the Type III task.

In the qualitative analysis of well-formed variation, we observed a considerable presence of variation at the level functional knowledge. Learners were using linguistic structures to express opinions that were not envisaged. This confirms that this task presents an indirect and broad relationship between input and response, and is therefore an open task in terms of thematic contents and linguistic contents.

From a pedagogical perspective, this is task seems to foster the a free choice in terms of thematic contents, as well as the use of specific exponents for language functions. We think this task qualifies as structured communication practice in Littlewood (2004)'s classification (see Section 7.2.2.4).

9.3 Response characteristics and NLP complexity

In this section we review some of the findings presented in the previous three sections to summarise the consequences that the performed analysis might have in the design and implementation of NLP-based feedback generation systems. We will centre the debate on the effects of length and variation as two of the dimensions that affect the complexity of the NLP task.

9.3.1 Response length

Table 9.12 shows the average length of the responses in terms of words for each activity. The differences in the length of the responses to the Type I (10.5 words) task with respect to the length of the responses to Type III and Type IV tasks (90.1 and 82.3 words), correspond with their typology in terms of expected response: the former task requires a limited production response, the other two an extended production response.

The higher standard deviations for the average length of the responses to tasks of Type III and IV in Table 9.12 support the argument that these activities tend to elicit responses of irregular length. Though this should be confirmed with statistically more significant numbers, we hypothesise a major difficulty in developing FSA-based

NLP resources for the analysis of texts with largely differing lengths. Tasks of Type III and IV seem to be activities that present a major challenge for NLP.

As a matter of fact, in Sections 8.4.1.1 and 8.4.3, we described how the modelling of the NLP resources to analyse limited production responses is less complex than that required to analyse extended production responses (cf. the automata in Figure 8.6, which suffice for the modelling of limited production responses, with the automata in Figures 8.14 and 8.15, to model a longer response). The differences between the these automata to model one type of response or the other make evident that the length of the responses is correlated with the size of the automata.

	# Resp.	Word avg.	SD
Act Type I	29	10.5	2.8
Act Type III	14	90.1	21.9
Act Type IV	9	82.3	32.1

Table 9.12: Response length average in words per activity type.

Another interesting aspect is that the standard deviation for the response length of responses to the Type IV task is more than 10 points higher than the standard deviation for the responses to the Type III task (32.1 in front of 21.9). This might be reflecting that the different relationships between input and response for these two activities: The Type III task presents a narrower and more direct relationship between input and response than the Type IV task. This can be used to draw conclusions affecting both the NLP design and the FLTL design.

9.3.2 Response variation

In line with Bailey and Meurers (2008), our hypothesis is that the more variation there is in the responses to a task the more difficult it will be to handle it with automatic analysis tools. In the following paragraphs we present the conclusions that we can draw from our analysis that can help us better describe this argument.

First of all, for the three types of tasks analysed, the ratio between the number of envisaged patterns used and the number of responses using them for each set of responses (5:14, 8:38, and 10:27) is always higher than the number of non-envisaged patterns used and the number of responses using them (14:15, 61:100, 25:35). That the proportion of envisaged patterns in learner response for the Type I task (13/16) is closer to that of the Type IV task (10:27) than to the Type III task (38/100), is, in our view, because the assessment strategy defined for the Type IV task focuses only on the linguistic contents, that is, the formal aspects – recall that this was decided in view of the difficulty foreseen in implementing an analysis strategy for an activity whose thematic contents were notably open. This supports the idea that assessing form is easier than assessing meaning.

Another interesting aspect about the proportions between envisaged patterns used in the learner responses and non-envisaged patterns used in the learner responses is that both seem to be worth the effort. On the one hand, design-based specifications provide the patterns to handle a percentage of the learner responses – certainly not

	Unexpected	Missing	Expected
Act Type I	0	2	98
Act Type III	6	12	82
Act Type IV	0	22	78

Table 9.13: Percentage of response element annotations indicating unexpected, missing or expected content.

easy to determine this percentage with the small response sets we used. On the other hand, when the number of responses, or response elements, is high enough a subset of patterns emerge that cover a good percentage of the responses, and this is true for the analysis of linguistic contents as well as for the analysis of thematic contents, independently of the type of task. This supports the viability of corpus-based development of ICALL systems including a careful cross-disciplinary design.

Second, the different analyses show that both well-formed variation and ill-formed variation occur frequently in learner texts. This supports NLP approaches to developing ICALL systems that focus strongly in the handling of errors. However, it supports also the line of research proposed in Bailey and Meurers (2009) that proposes the use of NLP techniques that allow for the analysis of unseen language structures by extending the coverage of the tagging and parsing strategies through properties of the language, not through an extension of specifications.

Third, for the tasks and the learner profiles in consideration, errors at the level of graphology and at the level of syntax are the most frequently seen. This supports the idea of applying (adapted) spell and grammar checking modules as part of the ICALL system, as a means to improve the performance of content analysis. In this respect, it is important to consider that semantic and textual coherence errors are the two types of errors that follow the other two. Since we are interested in making ICALL viable in learning contexts in which communicative language teaching is used, specific strategies for the handling of this type of error should be designed too.

Fourth, the amount of responses in which there is information missing, that is, responses in which part of the thematic contents is not included, seems to increase with response length and the broadness and the indirectness of the relationship between the input data and response. As reflected in Table 9.13, for the task requiring the shorter response there is a smaller percentage of response elements annotated as missing, that is, not included in the response. As for the other two tasks, the percentage of response elements missing is higher in the Type IV task, the more loosely restricted one, than in Type III task.

And, finally, as for the presence of unexpected thematic contents, they were only observed in the Type III task. It is interesting to see that no unexpected contents were found in the Type IV task, though the task itself is clearly less restrictive in these terms. This might reflect a trend in learners to resource to previous knowledge in tasks with longer texts, as foreseen by design in the Type IV task.

9.3.3 Section summary

To sum up, the qualitative and quantitative analysis presented suggests that both response length and the relationship between input and response have a role in the complexity of the NLP tools required to analyse learner responses. Response length and more restrictive tasks are more easily handled than longer, less restricted responses. However, a focus on the linguistic contents of the response, rather than on the thematic contents, tends to make the task's contents easier to predict. As for well-formed and ill-formed variation, they both occur frequently and in all tasks, and in fact well-formed variation is higher in tasks requiring longer and/or more loosely restricted responses.

As for the qualitative analysis of the Type IV task, it suggests that variation within response elements occurs not only when similar contents are expressed, but crucially also when similar communication acts (e.g., expressing an opinion) occur. This is critically different from what occurs when expressing more fixed, and short, communicative elements such as openings and closings in letters and emails, where the difference between the three tasks analysed is almost non-existent.

9.4 Learner data to improve the analysis strategy

From the NLP perspective, a straightforward application of such a quantitative and qualitative analysis is to improve and extend the precision and the recall of the resources and the feedback generation module. This could be framed within an iterative process of redefinition of the specifications and reimplementation of the corresponding modules and resources. Interestingly, this iterative approach is compatible with the workflow proposed by FLTL and CALL researchers for the development of learning materials.

In the following two sections we describe how from the analysed data two particular aspects of the NLP resources described in Chapter 8 could be improved.

9.4.1 Corpus-driven domain adaptation

As we saw in Section 9.2.3.1, almost one third of the responses using non-envisaged patterns to express an *Opinion* as part of the letter of opinion in the Type IV task included the verb *ser* [to be]. Most of this expressions can be interpreted as an opinion because of the adjectives and the topics used in the context. In (51), (52), (53), and (54), there is a pattern that can be derived from the sentences below, all of which are extracted from the analysed sets of responses. By adding the corresponding FSA recognition to the NLP analysis module, the assessment of the responses to this task would gain coverage of what can be considered an opinion in that context.

(51) X es práctico/conveniente.

a. La caja es muy práctica.
b. El envase es muy práctico.
c. El envase es muy conveniente.

(52) X es refrescante (para)

a. [Los Smint] Son muy refrescantes para después de comer o antes de una cita.

(53) X es imprescindible para

a. [Smint] es imprescindible para el buen aliento.

(54) X es (ADJ MODIFIER) caro

a. [Smint] es un poco demasiado caro.

However, in implementing the rules for the above patterns, a pattern such as the one underlying the sentence in (55) should be avoided. It has the verb to be (*era*, [was]), it has the adjective *posible* [possible], and the clause *hacerlos más grandes* [make them bigger]. So despite presenting the lexico-syntactic pattern X to be ADJ, the presence of *quería saber* [I wanted to know] is crucial to filter the sentence out as a positive opinion, and male rather a question. This is usually feasible in NLP-based strategies to be applied in a particular domain, as the domain of the task would be.

(55) Quería saber si era posible hacerlos más grandes.

9.4.2 Corpus-driven mal-rule approach

In the responses to the Type III task we found a couple of learners using the preposition *at* in the expression *work* at *the Marketing Department*, instead of *work* in *the Marketing Department*. As it happens to be the case, Spanish learners of English tend to mix up prepositions, particularly location prepositions, since there is not a clear correspondence between *at* and *in*, and *a* and *en*. So in order to provide this learner profile with specific feedback explaining the differences between the use of these two prepositions in the context, the appropriate recognition patterns could be implemented. This would imply a strategy based on three (or more) recognition patterns as reflected in Figure 9.1.

Note that the analysis proposed does not distinguish between the prepositions *in* and *for* in the sense that it assigns them the same analysis, *Prep-Ok*. Note also that prepositions other than *at* will be handled differently; this shows finer-grained feedback can be time-consuming in hand-crafted systems and, under certain circumstances, the response time of the system can be affected. All these variables have to be taken into account.

9.5 Chapter summary

In this chapter, we analysed a set of responses obtained from learners working with the materials presented in Chapter 7. We presented a rationale and a scheme for the annotation process to help us elucidate the extent to which learner responses included the thematic and linguistic contents expected according to RIF specifications. With this analysis we are able to evaluate the predictability of learner behaviour, as well as to confirm the need for corpus-driven approaches as a complement to careful design of ICALL tasks.

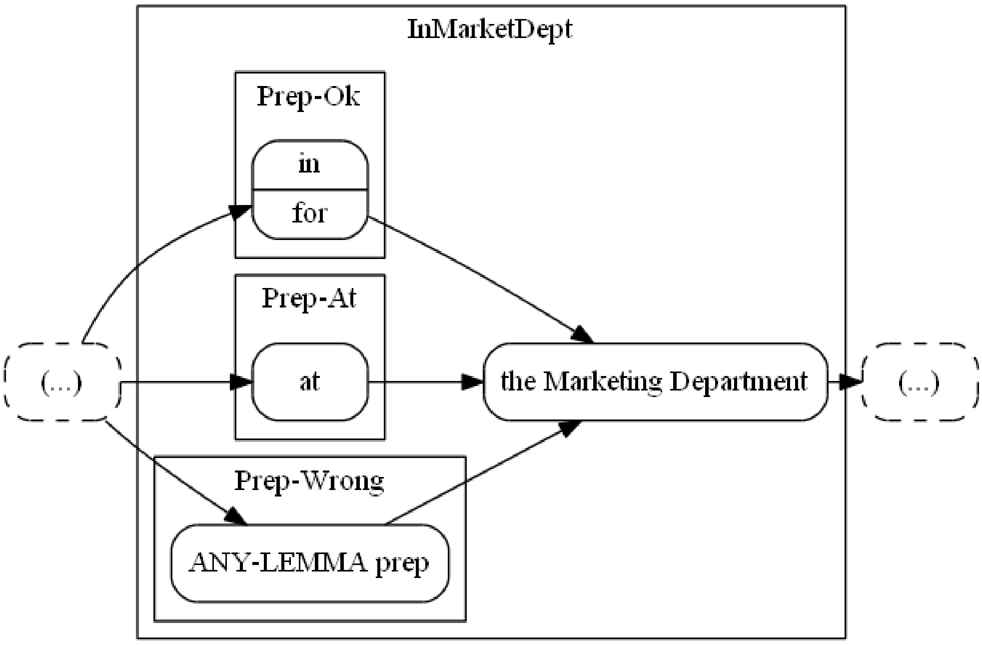


Figure 9.1: Graphs presenting the recognition paths required for a finer grained feedback to preposition errors in task-relevant language.

Our analysis confirms that tasks whose responses are longer and present a broader and less direct relationship between input and response pose a greater challenge in terms of NLP. This greater challenge is reflected in the presence of variation at levels of linguistic knowledge beyond the grammatical knowledge: sociolinguistics and textual knowledge. Our analysis confirms that well-formed and ill-formed variation occur in learner responses independently of their correctness and of the fact that they could be envisaged by design specifications. This speaks in favour of (i) necessarily complementing careful design with corpus-driven approaches and (ii) including strategies and modules for the analysis of variation in the NLP tools for the automatic analysis of learner language. Our analysis also confirmed the influence of task instructions and input data, or the teacher herself or himself, on learner performance, though this was not the focus of our study.

Finally, we show how learner corpora helps us adapt the NLP strategies to the specific thematic and linguistic contents of the domain emerging from the task. Moreover, the analysis of learner language helps increase the robustness of the NLP tools and the fined-grainedness of the generated feedback by improving mal-rules. All in all, our analysis confirmed the importance of anticipation as part of the design of ICALL materials, a process that will be inherently iterative. In our opinion, this process should include the trialling of the materials with learners as a means to increase the performance of the tools for the analysis of learner language, but also as a means to better assess the accomplishment of the pedagogical goals – a strategy totally in line with the approaches proposed by the FLTL and CALL research reviewed.

Part IV

Enabling teachers to author ICALL activities

Another issue is how an instructor could write a lesson without learning micro-planner or becoming a system builder. There seems to be a technical solution available already, though we have not implemented it in our prototype.

An Artificial Intelligence Approach to Language Instruction
Weischedel, Voge, and James (1978: p. 238)

[P]rescriptive designs which preclude options in the presentational and instructional formats, may not be received kindly by materials developers because they demand control not only over content, but also over the way the content is presented as well.

Computer-Assisted Language Learning
Context and Conceptualization
Michael Levy (1997: p. 19)

[T]he integration of these [out-of-class work] elements needs to be thoughtfully and coherently designed, often with the needs and resources of the individual learner in mind.

CALL Dimensions
Options and Issues in Computer-Assisted Language Learning
Levy and Stockwell (2006: p. 11-12)

Chapter 10

Customisation of an NLP-based feedback generation strategy

This chapter is the first step in facilitating the use of ICALL materials in real-world instruction settings, while aiming to preserve the control over the design of the materials in the teacher's hands.[1] While in Part III of the thesis we presented a methodology for the design of ICALL materials, in which a custom-tailored NLP-based feedback generation strategy was possible through the collaboration of FLTL and NLP experts, in this chapter we present a strategy to replace the NLP developer with a technology that will use teacher expertise to automatically generate part of the resources of the NLP-based assessment module.

Our strategy is based on two assumptions: First, that teachers, after having designed an ICALL task following the RIF, will be able to generate a set of gold-standard responses. Second, that teachers can use the set of gold-standard responses as a basis for making explicit thematic and linguistic contents of task responses in a way that task-specific NLP resources can be automatically generated. To do so, we propose a methodological and a technical solution to allow for the customisation of the NLP-based feedback generation architecture presented in Chapter 8.

We first introduce the context of application for the NLP-enhanced technology that we envisage to assist teachers in the authoring of ICALL materials. We present the roles and the activities that we expect teachers and learners to perform.

Second, we present the modifications to the architecture for the generation of NLP-based feedback presented in Chapter 8 allowing it to be customised on the basis of a set of gold-standard responses provided by the teacher. As we will see, the customisation of the NLP-based feedback generation architecture requires two further instruments. On the one side, we propose a formal specification language to express this set of expected responses, so that it can be automatically processed to obtain the underlying thematic and linguistic contents of the responses according to expectations. On the other side, we present a strategy for the generation of the

[1]The contents of this part is connected to the work carried out by me as part of the research team in the AutoLearn and the ICE^3 projects, which were projects by the Education, Audiovisual and Culture Executive Agency under the Lifelong Learning Programme (LLP 135693-LLP-1-2007-1-ES-KA3-KA3MP and 510653-LLP-1-2010-1-ES-COMENIUS-CMP). More information on both projects can be found at http://ice3.barcelonamedia.org/courses.

NLP resources that will feed the customisable modules in the NLP-based assessment architecture on the basis of the specifications provided.

Third, the chapter introduces a methodology through which we expect teachers to produce the specifications needed for the generation of the NLP resources. This methodology is closely related to the formal specification language we define. The response specification language is the interface between the teacher's knowledge of the task responses and the NLP-based strategy for the automatic customisation of the NLP-based feedback generation strategy. The response specification methodology for teachers is connected to pedagogical concepts and represents a significant simplification of a standard grammar writing formalism in NLP.

Finally, we present how the response specification methodology for teachers is applied to a particular item in an imaginary ICALL task. We describe how the concrete response specifications for a task are used to model correct responses and are expanded with linguistic patterns that model further correct and incorrect responses.

10.1 Context of application

For the design and implementation of a technology and the accompanying methodology for teachers to be able to author ICALL materials, we assume a set of characteristics and activities in the instruction context. This instruction context, our context of application, is shown in Figure 10.1. There are three main actions taking place, as well as two user profiles, the teacher and the learner; it also shows the functionalities required for the technology. The actions are numbered in a *a priori* chronological order: no. 1, activity design, no. 2, activity implementation, and no. 3 activity use. Actions, whose label is within circle and oval forms, are performed by teachers or learners as marked. The functionalities required from the technology are identified using square forms.

Here is a detailed description of the actions, roles and functionalities foreseen:

1. **Activity design** Teachers conceive activities according to learner needs and pedagogical goals; this process presupposes the use of the RIF, a methodological framework that allows for the specification of a set of gold-standard responses.

2. **Activity implementation** Teachers specify the expected responses and are able to automatically generate activity-specific NLP resources for the assessment of learner responses.

    (a) **Authoring the activity** Through a graphical interface teachers author the activity including instructions, input data, supporting references, and a set of expected responses.

    (b) **Generation of NLP resources** An NLP-based strategy takes as input the set of expected responses and automatically generates the NLP resources required for the automatic linguistic analysis of learners responses, as well as for the generation of feedback.

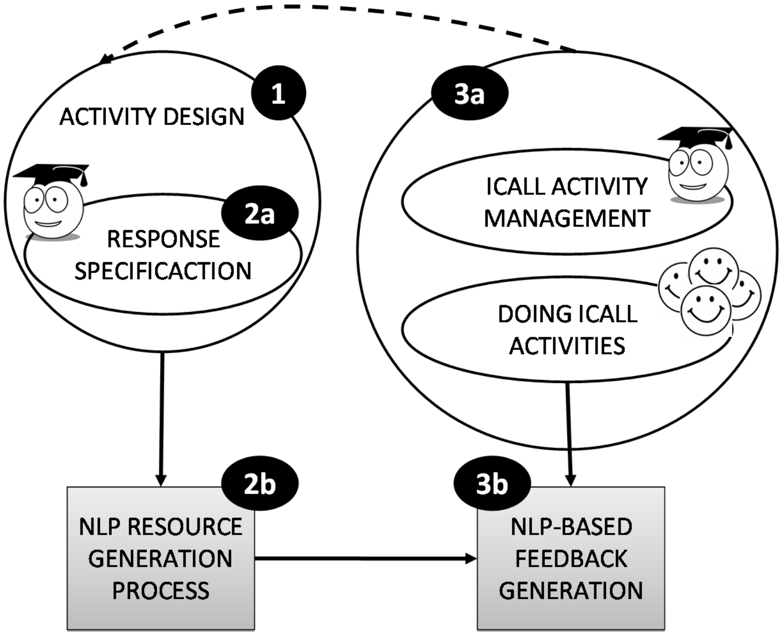


Figure 10.1: Context of use of an ICALL activity authoring and management tool including NLP as an enabling technology.

3. **Use of the activity** The ICALL activity is used in a setting where a blended learning approach is followed and where CALL materials are delivered through a virtual learning environment.

   (a) **Activity completion** Learners do the activity individually and at their own pace using a computer, in class or at home.

   (b) **Activity assessment** An NLP-enhanced technology provides learners with immediate automatic formative feedback to their responses.

A software solution to perform the above described actions and roles requires more than the implementation of a customisable NLP-based feedback generation strategy. Particularly, it involves a great deal of software design and programming, integration in a learning management system, design and development of graphical interfaces, specific teacher training materials, and so on. Though we acknowledge the importance of these different issues and, in fact, they were extensively investigated in the larger framework in which this research has been carried out, we strictly focus on the action *Activity implementation*, no. 2 in Figure 10.1. Our focus is to turn NLP into a so called enabling technology, that is, to use NLP as a means for teachers to achieve the autonomous authoring of ICALL materials without NLP training – as mechanical engineering allows us to drive cars without (almost) any notion of mechanics.

10.1.1 Formative feedback as a functionality

Since our goal is to provides learners with formative feedback on the basis of teacher specifications, before we introduce how we propose to achieve this, we establish what formative feedback actually will imply in this context. For us, formative feedback consists in providing feedback regarding:

- The presence of the expected contents in the learner response
- The presence of unexpected contents in the learner response
- The correct sequencing of the informative and linguistic elements by:
  - Checking for the completeness of the response in terms of topical knowledge
  - Checking for the correctness of the response in terms of linguistic knowledge
- Distinguishing between form and meaning errors, that is, linguistic knowledge errors and topical knowledge errors.
- Generating feedback messages including, when possible:
  - Localisation of the piece of text to which the feedback is referred to, unless it is a general message
  - Explanation of the highlighted phenomenon, distinguishing between warnings, errors and facts about which the virtual tutor is unsure.
  - Instructions to correct or improve the text, including a correction proposal if feasible

These functionalities are available in the feedback generation architecture presented in Chapter 8, which, as we said, is our starting point.

10.2 Customisable NLP-feedback assessment

We propose three different instruments that allow for the customisation of the NLP resources underlying an NLP-based feedback generation module on the basis of a set of expected responses. First, the actual feedback generation architecture has to be customisable: which modules and how. Second, there has to be a formal language for the specification of the set of gold-standard responses, one that requires a minimum structure in the specifications and allows for the distinction of thematic and linguistic contents expected in the response. Finally, there has to be an algorithm that on the basis of the specified responses models both correct and incorrect responses, containing well-formed and ill-formed structures.

10.2.1 A customisable architecture

To make customisable the NLP-based feedback generation, we propose to customise the domain-specific modules in the architecture presented in Section 8.3. In Figure 10.2 the domain-specific modules, the Information Extraction Modules and the Global Response Checker, are identified using a light-blue background colour. As we described in Chapter 8, these two components are respectively responsible for the task-specific linguistic analysis and the task-specific feedback generation, while the other modules continue providing general-purpose functionalities. The customisation of these two architecture components requires a response specification language and an expansion logic to automatically generates the finite-state rules (see Sections 6.3, 8.4, and 8.5).

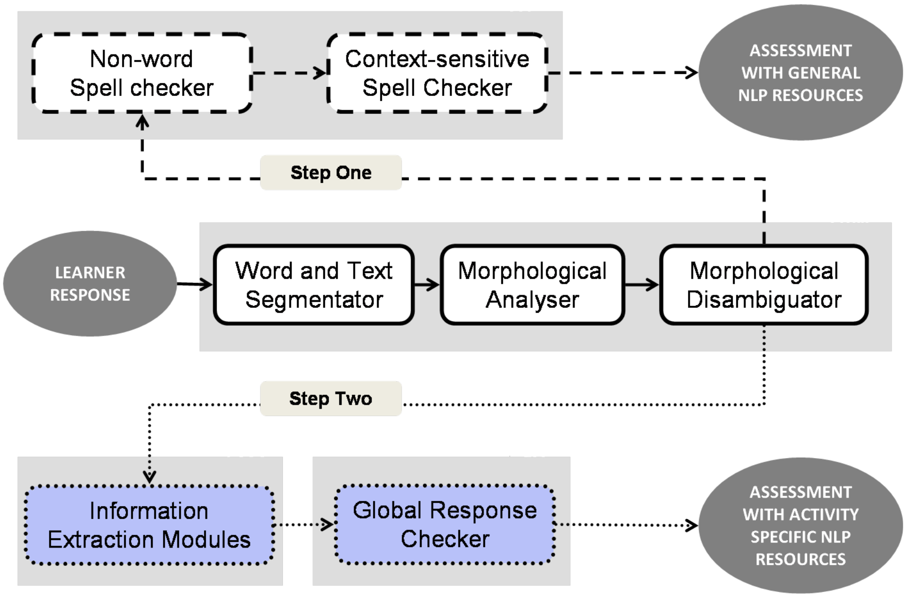


Figure 10.2: An customisable NLP architecture for the processing of responses to ICALL activities designed by FLTL practitioners.

The response specification language is a means for content developers to provide a set of *correct* responses. The set of correct responses contains, at least implicitly, the criteria for correctness of the ICALL task, as well as its thematic and linguistic contents in form of the language taught/learnt. In the customised solution, the Information Extraction Modules will use automatically generated NLP resources for the linguistic analysis of learner responses in order to detect specific text chunks corresponding to the response's expected contents. As for the Global Response Checker, it will be responsible for the evaluation of responses with respect to the task's criteria for correctness: the response's completeness, and its well-formedness.

10.2.2 Response Specification Language

The Response Specification Language (RSL) we propose relies notionally on the paradigmatic and syntagmatic properties of language, a concept introduced by Ferdinand de Saussure and extensively used in linguistics (Davies and Elder, 2004; Aarts and McMahon, 2006). Paradigmatic, or vertical, relations are those that can be established between linguistic objects that can be interchanged (Lyons, 1995: p. 126), for instance, at the syntactic level verb forms in a sentence can present a paradigmatic relation, as *plays* and *played* in the sentences *Mary plays/played the drums*. Similarly, letters would be in a paradigmatic relation within words, as *i* and *a* in *sit* and *sat*. Syntagmatic, or horizontal, relations are established between linguistic objects that can occur with one another in the same linguistic structure (Lyons, 1995: p. 126). For instance, agreement is a typical syntagmatic relation between subject and predicate in a sentence.

The RSL handles sets of responses as groups of linguistic objects that conveniently ordered build a set of sentences that correctly respond to a particular item in a FL learning activity. Responses are divided into components, which present paradigmatic and syntagmatic relations. Response components are used to build sets of sequences, where components can be added, omitted, substituted or reordered, though not all combinations will be allowed: only those building the set of correct responses.

10.2.2.1 Definition of the RSL

As schematised in Table 10.1, formally speaking the RSL consists of:

- **Responses**: A set of correct responses R consists of a list of Response Component Sequences; each sequence corresponds to one or a group of sentences that happen to be a correct response for the activity item.

- **Response Component Sequences**: A RCS is a list of ordered Variants: each Variant belongs to a different Response Component, that is, Variants are in syntagmatic relation with respect to Variants in other Response Components as part of and RCS.

- **Response Components**: A RC is a set of variants: only one of the Variants in a RC can be part of a RCS, that is, Variants are in paradigmatic relation within a response component.

- **Variants**: A variant V which is a set of Strings: each string is in paradigmatic relation with other strings in the same variant.

- **String**: A string S consists of ordered sequences of tokens.

- **Tokens**: A token t is a word or a textual symbol (punctuation, figures) belonging to a natural language – the foreign language.

- **Optionality operator**: The optionality operator ? is used to indicate that a variant in a RCS is optional.

- **Disjunction operator**: The disjunction operator | is used to indicate that one and only one in a list of variants within a RCS is compulsory.

Term	Definition
Responses	$\{RCS_1, RCS_2, ..., RCS_n\}$
Response Component Sequences	$< V_{i_{RC1}}, V_{j_{RC2}}, ..., V_{n_{RCm}} >$
Response Components	$\{V_1, V_2, ..., V_n\}$
Variants	$\{S_1, S_2, ..., S_n\}$
Strings	$< t_1, t_2, ..., t_n >$
Tokens	Words or symbols
Optionality operator	?
Disjunction operator	$\vert$

Table 10.1: Formal definition of the Response Specification Language.

10.2.2.2 RSL-compliant representation of expected responses

Figure 10.3 exemplifies an abstract representation of the specification of a set expected responses for an imaginary question. The imaginary response consisting of three response components are: A, B, and C. As shown in Figure 10.3a, each RC includes a list of variants: RC A has only one variant, A1, RC B has two variants, B1 and B2, and RC C has three variants, C1, C2 and C3.

Figure 10.3b represents the two possible RCS for this imaginary activity (for the purpose of the explanation). In both of them variant A1 is optional, marked with an interrogation mark (?). RC B1 and B2 are compulsory but excluding: If one of the linguistic objects in B1 is used, then this can only be followed by one of the linguistic objects in C1 or C3. If the selected linguistic object belongs to B2, then the following one has to belong to either C2 or C3. The optionality between C1 and C3 in the first RCS and the one between C2 and C3 in the second RCS is marked with a vertical bar (|).

10.2.2.3 Pedagogical and linguistic notions underlying the RSL

The RSL does not include explicit information in terms of pedagogical criteria, neither in terms of thematic or linguistic contents. However, there is a crucial relationship between some of the elements of the RSL and the two tasks at hand: the linguistic analysis of learner responses and the evaluation of the responses in terms of criteria for correctness. Table 10.2 reflects these relationships.

RSL concept	Related to
Response Component Sequences	Linguistically determined combinations of RCs
Response Components	Activity's criteria for correctness
Variants	Linguistic realisations of an RC

Table 10.2: Pedagogically and linguistically relevant concepts of the Response Specification Language.

From a pedagogical perspective, Response Components (RCs) are relevant in terms of thematic contents. RCs correlate with the concepts expected in the response;

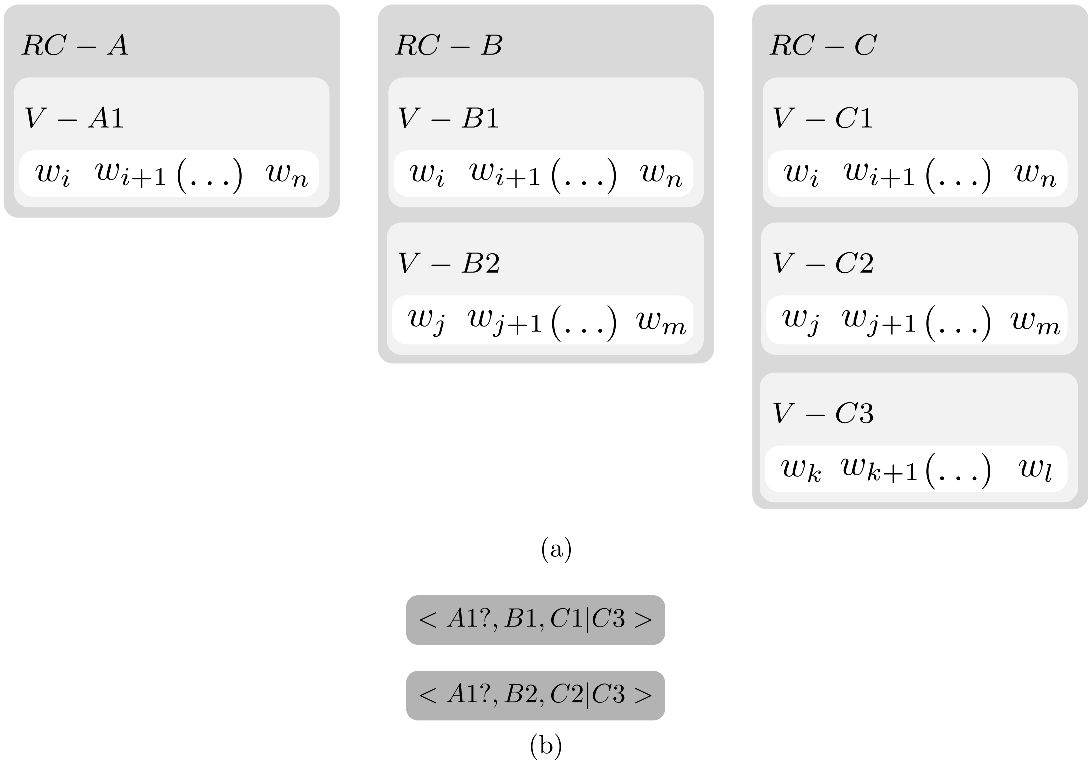


Figure 10.3: Abstract list of Response Components and the corresponding RC Sequences.

they are minimal units of information corresponding to entities and/or relations expected in the response. As for Variants (V), they are linguistic realisations of the pedagogically relevant concepts, that is, of the different RCs. In other words, Variants are subordinated to RCs, in the sense that each Variant is a different way of expressing that same concept. Thus, Variants correlate with the linguistic contents expected in the response.

Finally, Response Component Sequences (RCS) are the bridge between this thematic and the linguistic contents expected in the response. RCSs determine the different RCs required as well as the restrictions applicable according to the linguistic characteristics of the Variants chosen. RCSs are the representation of the different ways of complying with the thematic and linguistic criteria for correctness.

10.2.3 Customisable modelling of correct and incorrect responses

The feedback generation strategy that we propose is based on the modelling of NLP-based response evaluation strategies that presuppose the modelling of responses that match exactly with the provided specifications, and the modelling of responses that partially match with the provided specifications with the proviso that the deviations with respect to specifications can be used to provide pedagogically motivated feedback.

To guarantee the modelling of correct and incorrect responses on the basis of customisable NLP resources, we apply two different types of operations on the responses in RSL format. These two types of operations will be used to generate the corresponding finite-state recognition patterns. First, we foresee an automatic linguistic analysis of the linguistic elements of each of the Variants, that is, of each of the strings in a Variant. This linguistic analysis will make available three different levels of linguistic description: words, lemmas, and POS tags, including a minimum amount of mophosyntactic features. Second, we foresee the use of transformation operations to be performed on recognition patterns to model phenomena such as the presence, the absence or the transposition of nodes with respect to the recognition paths derived from the specified correct responses. These two types of operations will limit the variation handled by such a customisable strategy for the generation of NLP-based feedback.

10.2.3.1 Modelling exact matching responses

The modelling of exact matching responses results from generating the corresponding sets of finite-state rules at the word level. Variant specifications as in Figure 10.3 result into rules for the Information Extraction Modules, where each string results in recognition paths labelled with the corresponding variant identifier. Figure 10.4a shows the recognition paths that the variant specifications in Figure 10.3a would yield. Each string of tokens consists of a series of nodes that uses the word analysis level using the notation w_ι, where w stands for a word and $_\iota$ for its index number.

Response Component Sequences as in Figure 10.3b result into rules to be included in the Global Response Checker. Note that the nodes in the recognition paths in this module use variant labels (A1, B1, B2, ...); the recognition of the corresponding nodes in the corresponding order in a particular response allows for the checking of the expected thematic contents in the response, expressed using the expected linguistic contents.

Optionality and disjunction, as in Figure 10.3b, are converted into specific recognition paths. For instance, variant V-A1 is made optional by allowing variants V-B1 and V-B2 to be starting nodes, marked with a dashed border – the dotted border in the C variants indicates that they are ending nodes. Thus, Figure 10.3b reflects four different possibilities of starting a correct response by combining the variant nodes $A1$, $B1$ and $B2$: $A1 \rightarrow B1$, $A1 \rightarrow B2$, $B1$, and $B2$.

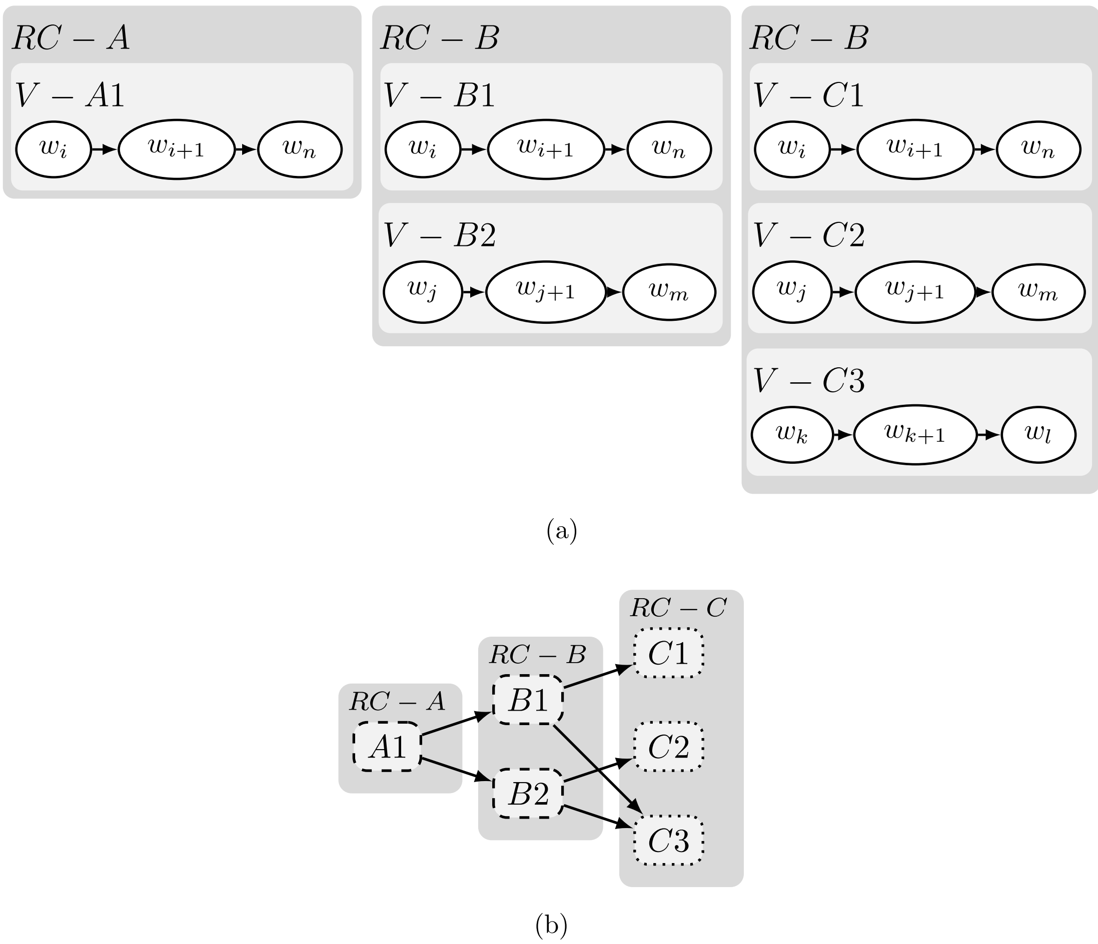


Figure 10.4: Customised FSA recognition paths derived from the RSL specifications in Figure 10.3.

10.2.3.2 Modelling partial matching responses

The modelling of partial matching responses results from exploiting the linguistic information obtained through the POS tagging of specified correct responses combined with the application of one or more transformation operations to the FSA rule that represents the correct specified version of a response. Thus, when a learner provides a response, if the text does not match with any of the recognition paths corresponding to the exact matching responses, different options will be tried. If a learner response does not match at the word level, but it matches totally or partially at the lemma level, then the linguistic differences between the expected and the elicited response can be computed and an informed analysis can be generated. Similarly, POS tag information can be used to match for further kinds of deviations.

Whether a feedback message is generated or not should depend on the feedback generation strategy. For the purpose of this thesis, every difference between a learner response and the expected response(s) correspond with a feedback message, though not all feedback messages, as we will see, are error messages. Section 10.3 provides concrete examples of the kinds of rules that can be generated.

10.2.3.3 Transformation operations

The transformation operations that we use for the expansion of the expected correct responses are based on a taxonomy of transformations originally used to analyse errors in the surface structure of text produced by learners (Dulay et al., 1982: p. 150). In the following sections we describe the types of transformations foreseen.

Transformation operations can be applied to FSA recognition paths both in the IEM and the GRC components. However, not all of the transformation operations or all of the linguistic levels of description are equally interesting or useful. Linguistic and pedagogical insights can help reduce or determine the nature and the amount of expanded recognition paths. In all the examples that follow we assume that the nodes contain word, lemma and POS tag information associated to the Variant tokens, as well as internal labels that correlate with RCs that can be used to determine the appropriate order of Variants according to the specified RCS.

Omission

Omission of path nodes, or associated information, is foreseen in three different positions in the recognition path: at the beginning, in the middle or at the end of the recognition path. This yields to transformations as those reflected in Figure 10.5, which reflects the different possibilities assuming a four-node variant as the one reflected in the recognition path in Figure 10.5a. Figure 10.5b shows how the initial node in the recognition path can be made optional through allowing the first middle node to be a starting node. Figure 10.5c shows how the middle nodes, one or both of them, can be made optional by allowing transitions that skip them. Figure 10.5d shows how the end node can be made optional by allowing the second middle node to be a final node.

Omission is only applied at the level of whole words, since the omission of one of the elements of the word, for instance, the lack of the morpheme for plural in a noun

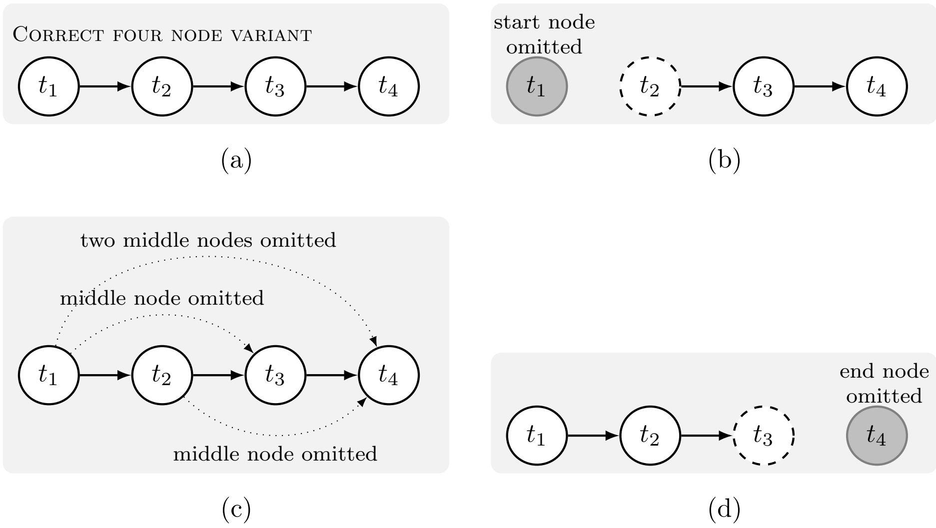


Figure 10.5: Expansion of RSL-based response patterns using omission transformation operations.

is handled as a substitution of one word form with another word form that has the same lemma, but different POS tag information. In Section 10.4.2.1 we show how the omission of several words is handled following a so-called bag-of-words approach.

Substitution

Transformation operations based on substitution are performed at the beginning, in the middle and at the end of the recognition path. Figure 10.6 shows the application of such an operation to the "correct" four-node variant in Figure 10.6a. Figure 10.6b shows a recognition path where the initial word was replaced with an unexpected one, marked as *unk*nown in the node path. The rest of the recognition path remains intact. Similarly, Figure 10.6d shows the recognition path where the node that replaced is the final node. Finally, Figure 10.6c shows the replacement of one of the middle nodes or both of them with an unknown word.

Addition

As shown in Figure 10.7, addition operations are also performed at the beginning, in the middle and at the end of the recognition path. Figure 10.7b shows a recognition path where there is an additional starting token. Similarly, Figure 10.7d shows a recognition path where the final node is converted into a middle node by accepting and additional node containing unexpected information after it.

Transformations in the middle nodes are those that generate more possibilities: Figure 10.7c shows the possibility that additional nodes are inserted between the first and second node, the second and the third and the third and the fourth. The figure also presents a looping arrow in the additional nodes so that more than one unexpected token can be recognised.

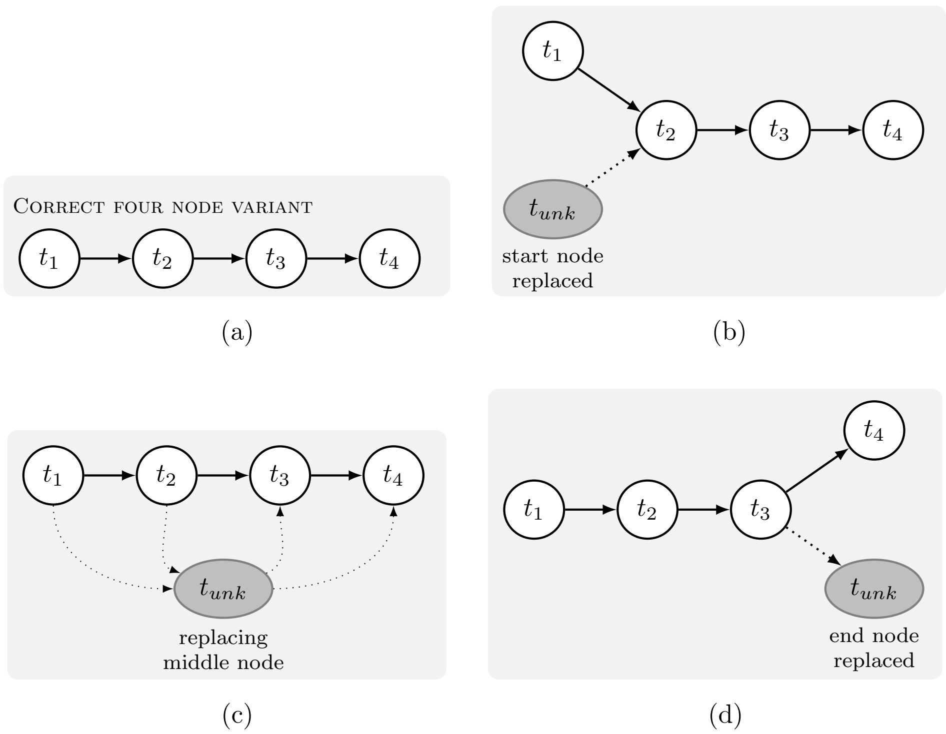


Figure 10.6: Expansion of RSL-based response patterns using substitution transformation operations.

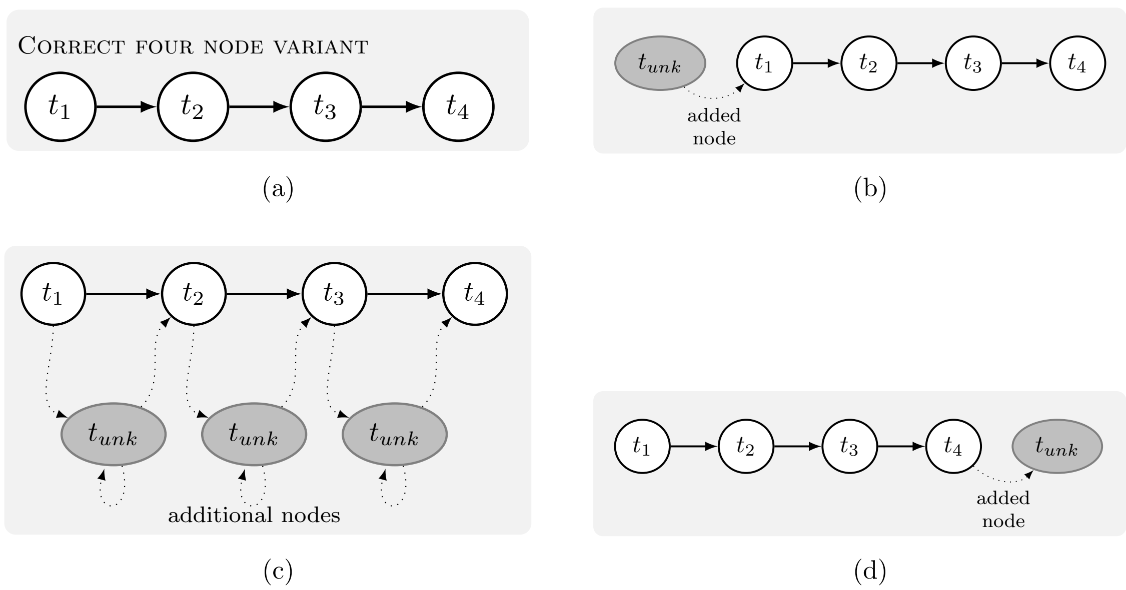


Figure 10.7: Expansion of RSL-based response patterns using addition transformation operations.

Reordering

Figure 10.8 shows the expansion of a correct four node recognition path using reordering as a transformation operation. For the sake of simplicity, we assume that only adjacent elements can be re-ordered. Figure 10.8b shows a reordering of the nodes in the initial part of the recognition path, where the first node can be recognised after the second. Figure 10.8c shows this reordering for the middle nodes, where the second node can be recognised after the third. And, finally, Figure 10.8d shows the result of reordering the third and the fourth node.

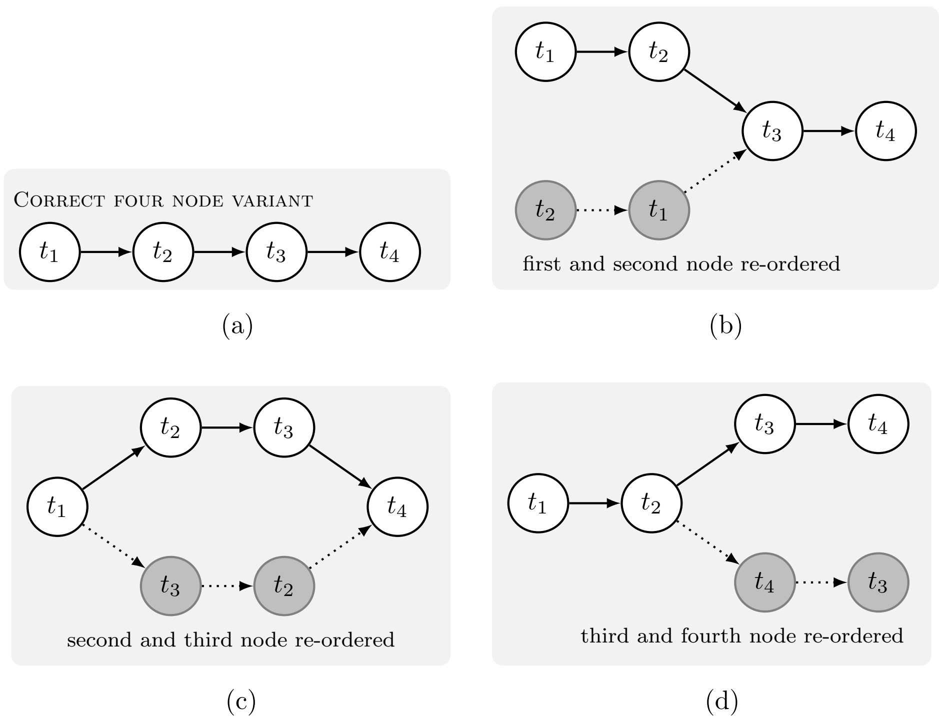
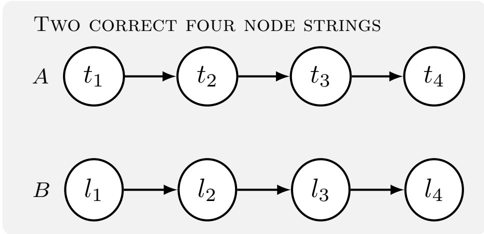


Figure 10.8: Expansion of RSL-based response patterns using reordering as a transformation operation.

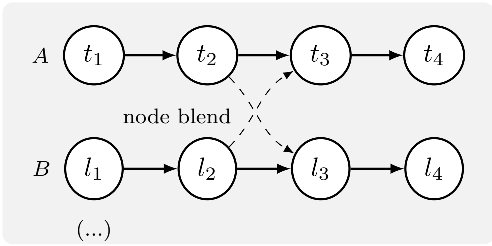
Blending

Figure 10.9 shows the expansion of two correct four node recognition paths using blending as a transformation operation. Note this transformation operation requires the existence of two different recognitions paths, otherwise it would be a reordering. Figure 10.9a shows the two correct four node strings, and Figure 10.9b shows the possibility of starting a sequence with $t_1 \rightarrow t_2$ and ending it with $l_3 \rightarrow l_4$, while it also allows starting it with $l_1 \rightarrow l_2$ and ending it with $t_3 \rightarrow t_4$.

We generate blending for elements that are in paradigmatic relationship, that is, t_1 can be replaced with l_5, because both of them are in the initial position, but it cannot be replaced with other elements in String B. The three dots below String B in Figure 10.9b indicate that through permutation of the four nodes in each of the strings other blending patterns could be generated, up to 14.



(a)



(b)

Figure 10.9: Expansion of RSL-based response patterns using blending as a transformation operation.

10.3 A methodology for teachers to author ICALL materials

In addition to equipping teachers with the strategy for the automatic generation of NLP-based response evaluation functionalities, they need a methodology to characterise the tasks pedagogically and linguistically. This methodology should help them (i) to decide whether a particular FL learning activity is suitable for being evaluated using NLP strategies, and (ii) to specify a set of predictable responses and evaluation criteria for that activity.

10.3.1 FL learning activities that suit NLP

With the Task Analysis Framework and the Response Interpretation Framework (see Sections 7.1 and 7.2), we can (i) characterise thematically and linguistically the responses to an ICALL task, (ii) establish the criteria for correctness, the type of response and the relationship between input and response, and (iii) generate a list of gold-standard responses. As shown in Chapter 9, analysing learner responses to particular FL learning activities can inform us about the complexity and feasibility of the activity in terms of linguistic variation. On the basis of this kind of analysis, teachers could consider the appropriateness for a FL learning activity to become an ICALL task. However, further research is required to convert those empirical findings into practical numerical rates or checklists to help teachers determine the feasibility of a particular FL learning activity.

For the purpose of this research, an FL learning activity will be suitable for being handled with NLP-based strategies if it is a Type I task (see Section 7.2.2.1). Type

I tasks are FL learning activities whose relationship between input and response is direct, and narrow in scope. In this type of task, the expected responses depend considerably on the input data.

10.3.2 ReSS: Response Specification Scheme

The Response Specification Scheme (ReSS) is the process that teachers follow to produce specifications that correspond to the pedagogical design of the materials and that can be used for the generation of the customised NLP resources. The application of the ReSS assumes that the questions for a given activity were characterised with RIF, and it requires content designers to divide responses into smaller parts and determine the order(s) in which these smaller parts combine to yield correct responses. The resulting specifications should be convertible into the RSL format (see Section 10.2).

Assuming a set of gold-standard responses derived from the RIF, we propose to follow this procedure:

1. Identify Response Components in the responses of the RIF-based list
2. Classify response fragments in each sentence into one of the possible Response Components, that is, identification of strings to be assigned to Variants
3. Identify combinatorial restrictions between Variants
4. Identify optional or alternative elements within a response fragment
5. Order Variants to generate the lists of correct Response Component Sequences

10.3.2.1 RIF-based characterisation of an activity

To exemplify the use of the ReSS for a particular ICALL task, we apply it to a fictive audiovisual comprehension activity in which learners are required to watch the film *E.T. The Extra-Terrestrial* and then answer the question *How did E.T. learn to speak English?*. By applying the RIF, we obtain the criteria for correctness and the corresponding list of gold-standard responses (see the complete TAF and RIF analysis in Appendix F).

The criteria for correctness for this question are:

- To include in the response a reference to the entities:
  - E.T.
  - Sesame Street (TV programme, television)
  - words
  - Gertie or Elliot's sister (OPT).
- To include also a reference to the relations:
  - word repeating, that is, E.T. repeated words

  - E.T. heard or listened to words
  - E.T. (or Gertie) watched Sesame Street
  - E.T. learnt English. (OPT)

- Responses have to be in the past tense, that is, historical present is not allowed.
- Responses containing an adverbial sentence expressing *how* E.T. learned English suffice – i.e., full sentences are not compulsory.
- Minor graphical errors such as punctuation or capitalisation errors are allowed.

The list of RIF-based gold-standard responses in (56) is obtained from the above listed criteria.

(56) a. By repeating the words it heard watching Sesame Street.
b. E.T. learnt to speak it by repeating what it heard watching Sesame Street.
c. E.T. learnt to speak English by repeating what it heard watching Sesame Street.
d. By repeating the words that it listened to while Elliot's little sister watched Sesame Street.
e. By repeating what Elliot's little sister said in response to her watching Sesame Street.

10.3.2.2 Applying the ReSS to a set of expected responses

Using the ReSS we can convert the set of gold-standard responses into RSL-format structure.

Step 1: Identify Response Components in the response

In accordance with the criteria for correctness, we identify four different response components for this activity item, as shown in Figure 10.10. All response components are at the level of topical knowledge: They correlate with the relations specified in the thematic contents of the RIF analysis (see Annex F) and the criteria for correctness.

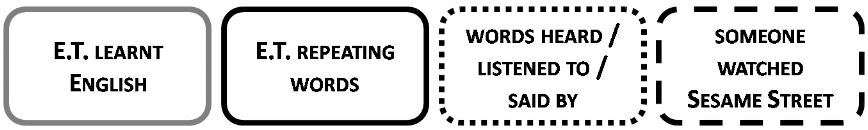


Figure 10.10: Response Components of the response to the E.T. comprehension activity.

We can see that RCs are pedagogically relevant because of its connection to the criteria for correctness: If all the expected thematic contents are present in a response – in one of the expected forms and orders – then it complies with the criteria for correctness.

Step 2: Classify response fragments in each sentence to one of the possible Response Components

Figure 10.10 uses different border styles in the boxes that identify each response component, and these styles are also used in Figure 10.11 to identify the text chunks in the sentences in (56) that correspond to each response component.

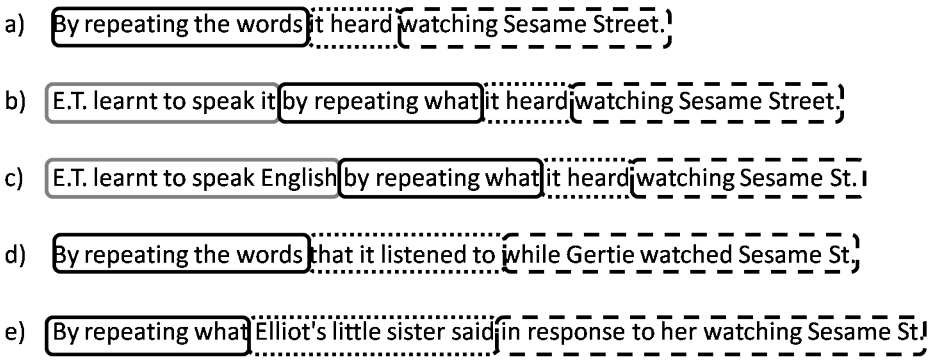


Figure 10.11: Identification and classification of the Response Components.

Step 3: Identify combinatorial restrictions between Variants

Response components with only one variant Figure 10.11 shows the text chunks in (57) that were classified as elements of the response component E.T. LEARNT ENGLISH, using a grey plain border. These text chunks only present combinatorial restrictions with those of the response components that follow them, not with those that precede them, because they are only used at the beginning of a sequence.

The sentences in (56b) and (56c) are equally correct as a response to the activity's question, and they would still be independent of the text chunk used among the two text chunks in (57). There is no reason to identify more than one variant in this response component, since all of its text chunks are equally compatible.

(57) a. E.T. learnt to speak it [...]
b. E.T. learnt to speak English [...]

Response components with more than one variant In contrast, the text chunks of the response component E.T. REPEATING WORDS, marked with a black plain border, present a different behaviour, as reflected in (58). Out of the three text chunks in it, two of them can be used at the beginning of the sentence, (58a) and (58b), but the third one can only be used in the middle of the sentence, (58c).

(58) a. By repeating the words [...]
b. By repeating what [...]
c. [...] by repeating what [...]

This difference argues for grouping (58a) and (58b) into one of the variants of the response component, while (58c) will be the only text chunk of a different variant. If we take into account that (58a) and (58c) only differ in the case of the first word, we could group the three text chunks in (58) in the same variant.

As shown in the combinations in (59) and (60), the three text chunks in (58) are not equally compatible with the text chunks representing the RC WORDS LISTENED TO OR SAID BY. Therefore, the initial grouping of the strings of the RC E.T. REPEATING WORDS should be revisited. The combinations in (59) are literally extracted from the list of gold-standard responses, while the ones in (60) result from re-using the text chunks originally not combined – texts corresponding to the RC WORDS LISTENED TO OR SAID BY are in italics.

(59) a. By repeating the words *it heard* [...]
b. By repeating the words *that it listened to* [...]
c. [...] by repeating what *it heard* [...]
d. By repeating what *Elliot's sister said* [...]

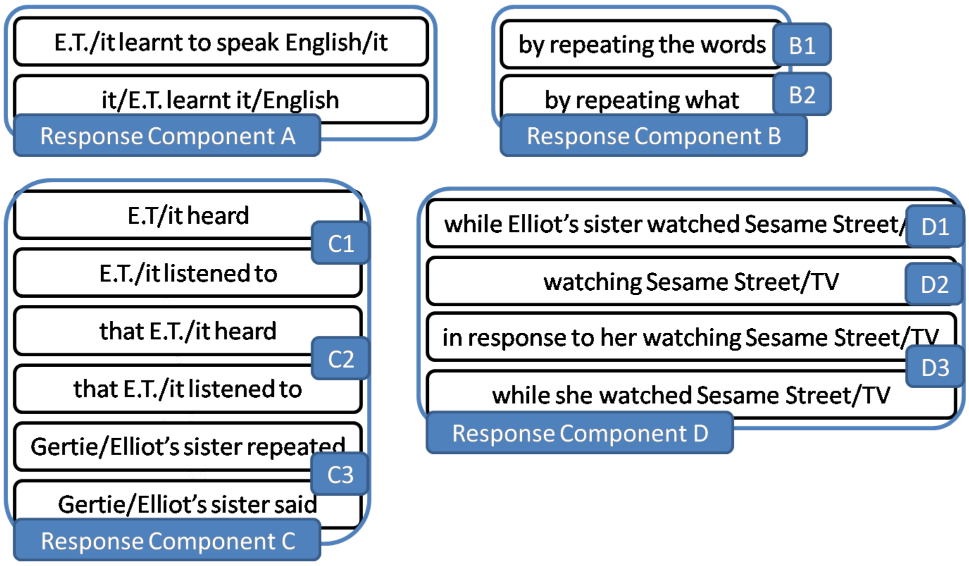
(60) a. By repeating the words *Elliot's sister said* [...]
b. [...] by repeating what *Elliot's sister said* [...]
c. * [...] by repeating what *that it listened to* [...]
d. By repeating what *it heard* [...]
e. * By repeating what *that it listened to* [...]

As shown in (60c) and (60e), when combining the response fragment *by repeating what* with a text chunk starting with *that* we obtain a linguistically ill-formed sentence. If the object of *repeating* is *the words* the following linguistic element being a subordinate clause might start with *that* or not. But if the object of *repeating* is *what*, then the following linguistic element cannot start with *that*, because *what* and *that* are not syntagmatically compatible. They are both performing the role of a relative pronoun that connects the preceding and the following clause, and the English language system does not accept this structure.

By systematically applying steps 2 and 3 to the gold-standard responses, we can produce a ReSS graph as the one in Figure 10.12, which represents the response components of the question under consideration. The four of them are identified with the letters A, B, C and D, corresponding to the response components E.T. LEARNT ENGLISH, E.T. REPEATING WORDS, WORDS HEARD / LISTENED TO / SAID BY, and SOMEONE WATCHED SESAME STREET. Some of the response components consist of only one variant, such as response component A, while others consist of two or more variants, such as B, C and D. Variants are named using a letter and a numerical index: B1, B2, for B, C1, C2, C3, for C, and D1, D2 and D3, for D.

Step 4: Identify optional or alternative elements within a response fragment

As reflected in Figure 10.12a, the result of identifying optional or alternative elements in text chunks is marked with a slash (/) to show optional or alternative linguistic



(a)

RCS 1	$< A1?, B1, C1, D1 >$	*RCS 7*	$< A1?, B2, C1, D1 >$
RCS 2	$< A1?, B1, C1, D2 >$	*RCS 8*	$< A1?, B2, C1, D2 >$
RCS 3	$< A1?, B1, C2, D1 >$	*RCS 9*	$< A1?, B2, C3, D2 >$
RCS 4	$< A1?, B1, C2, D2 >$	*RCS 10*	$< A1?, B2, C3, D3 >$
RCS 5	$< A1?, B1, C3, D2 >$		
RCS 6	$< A1?, B1, C3, D3 >$		

(b)

Figure 10.12: Graphical representation of the Response Components and the correct RC Sequences for the E.T. comprehension activity.

objects. For instance, response component A contains, among others, the text chunks *E.T. learnt English*, *E.T. learnt it*, or *It learnt it*, where *E.T.* and *It* or *English* and *it* are alternative elements in this component.

Step 5: Order Variants to generate the lists of correct Response Component Sequences

Figure 10.12b is the result of ordering response component variants taking into account which syntagmatic relations, i.e., order, and which are to be included in a correct response according to the criteria for correctness. Since Variant A1 is always optional, so a total of 20 different recognition paths are possible.

The different recognition paths in Figure 10.12b show that from B1, any of the variants in C can be reached, but from B2 only C1 and C3 can. C2 cannot be reached from B3 because of the *what-that* incompatibility. The figure shows that C1 and C2 can be followed by both D1 and D2, but not by D3. Since in D3 there are anaphoric references to a feminine singular individual, which is not present in C1 and C2, they would yield non-interpretable sentences such as the one in (61).

(61) #*By repeating the words it heard while she watched Sesame Street.*

In contrast, from C3, D2 and D3 can be reached, but not D1. Combining D1 and C3 would yield the sentence in (62), a sentence that is not only stylistically poor, but pragmatically dubious: It uses a proper name twice in two different clauses in the same sentence. It might be acceptable if one needs to mark it specifically, for instance, to state that this person was doing both actions, repeating or saying words, in a context where it was specified that one of either actions was done by someone else, but this is clearly not the case in the context of the activity.

(62) #*By repeating the words Elliot's sister repeated while Elliot's sister watched Sesame Street.*

10.4 Generating activity-specific NLP resources

This section exemplifies the generation of the recognition paths out of the specifications in Figure 10.12 with the expansion strategies described in Section 10.2.3.

10.4.1 Exact matching responses

Figure 10.13 exemplifies the recognition patterns to analyse the corresponding variants at the level of Information Extraction Modules in the learner response. These are two recognition paths at the word level for the analysis of the response component E.T. REPEATS WORDS, with its two variants, B1 and B2. Similar recognition paths would be generated for each response component: E.T. LEARNT ENGLISH, E.T. REPEATING WORDS, WORDS HEARD / LISTENED TO / SAID BY, and SOMEONE WATCHED SESAME STREET. In our example in Figure 10.13, there is only one string, or recognition pattern, per variant, but for some of the variants, like C1, C2, C3 and D3, the corresponding amount of recognition paths would be generated.

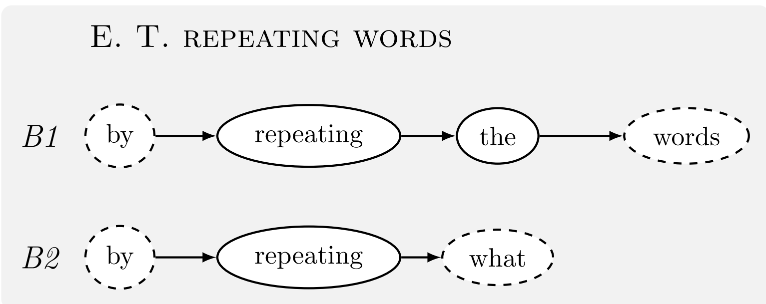


Figure 10.13: FSA recognition paths generated for the strings of Variants B1 and B2 in the RC E.T. REPEATING WORDS.

Figure 10.14 shows the recognition path that would be part of the Global Response Checker to check for the presence and correct ordering of all the expected response components. It is based on the RCSs defined in Figure 10.12b.

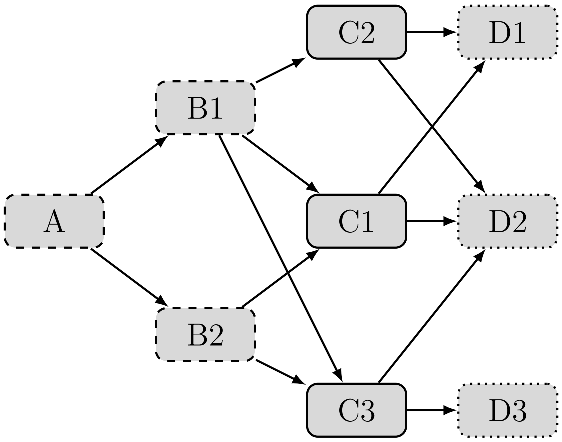


Figure 10.14: FSA recognition paths generated for the checking of exact matching RC sequences for the E.T. activity.

Note all the recognition paths start with either A, the variant of the response component E.T. LEARNT ENGLISH, or B1 or B2, the two possible variants of the response component E.T. REPEATING WORDS. This is marked by the dashed borders of the corresponding nodes: The three of them are valid starting nodes and this makes variant A optional. If B1 is reached, any of the variants in RC C, WORDS HEARD / LISTENED TO / SAID BY, can follow. However, if B2 is reached, only variants C1 or C3, can follow. After C2, D1 or D2 can follow, and after C1 any of the D response components can follow. After C3, D2 or D3 are accepted, but D1 is not. D variants are ending nodes belonging to the component SOMEONE WATCHED SESAME STREET – with a dotted border.

10.4.2 Pre-envisaging deviating responses

The handling of deviating responses is not necessarily related to the handling of incorrect responses: There might be a range of variations derived from the responses specified using the ReSS that are correct. On the basis of a POS-based linguistic analysis the general transformation operations presented in Section 10.2.3.3 can be used to expand the ReSS-specified expected responses.

10.4.2.1 Variation derived from omission

Omission is foreseen within strings in variants, and then the expanded recognition paths will be included to the IE modules, or within RCS, and then the expanded paths are included in the GRC.

Omission within variant strings

We foresee two types of omission at the level of strings: The first one is the omission of certain types of function words, like prepositions, determiners or conjunctions. The second type is related to the absence of relevant vocabulary words, often called content words, which correspond to adjectives, nouns and verbs. Both types of variation are exemplified in Figure 10.15.

Figure 10.15a reflects three different recognition paths, labelled B1.1, B1.2 and B1.3, obtained by expanding the original string *by repeating the words*, labelled B1. B1.1 and B1.2 reflect the omission of either the preposition or the determiner. B1.3 reflects the possibility to have both omitted. Thus, the response component would be recognised independent that the preposition *by* or the determiner *the* are present in the learner response. The different variant labels identify the response component assigned, as well as the corresponding deviation. The empty sets in the different paths indicate nodes omitted with respect to the original specifications; the starting nodes, indicated with dashed borders, are modified if required.

Figure 10.15b shows a path that recognises sequences of expected content words in the corresponding response component. For B1, a two-word sequence built with *repeating* and *words* will suffice; for B2, the accepted words are *repeating* and *what*. Note these statements, these recognition paths, are rather vague, since they do not require any order in the detection of the sequence.

Note both expansion strategies are only possible if the ReSS specified strings of words are POS tagged and the corresponding POS information is available for both function and content words.

Omission within response component sequences

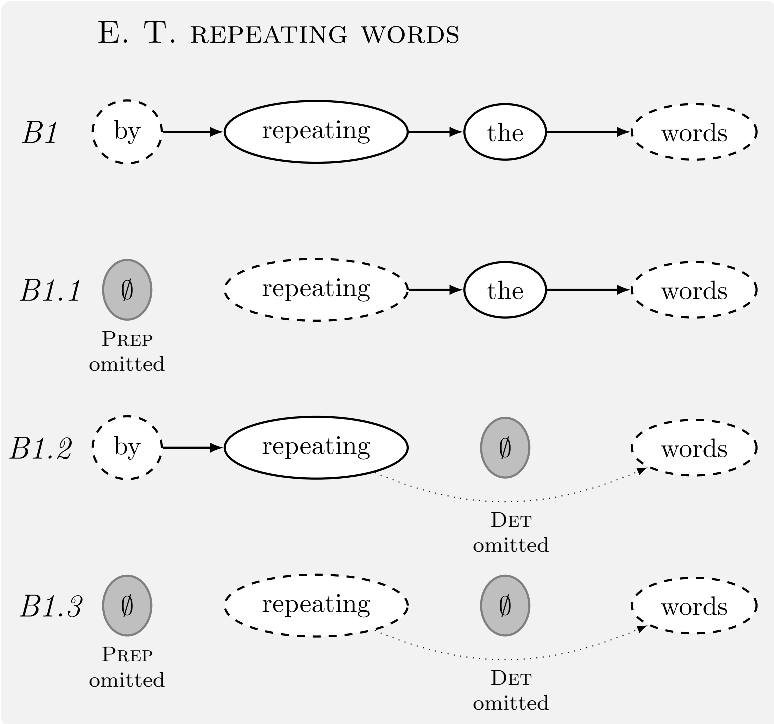
Figure 10.16 reflects omission at the level of response component sequences for one of the response sequences in Figure 10.14, namely the path A $\rightarrow$ B1 $\rightarrow$ C1 $\rightarrow$ D1. The figure exemplifies three cases in which one of the response components is omitted: RC1 is the expected and specified sequence, RC1.1 is the sequence where B1 is omitted, RC 1.2 is the one where C1 is omitted, and RC 1.3 is the one where D1 is omitted. The omission of A is not generated because it would yield a valid RC sequence, since A as a response component is optional.

For each of the RC sequences derived from Figure 10.12b, a similar process of pattern expansion could be followed. The parentheses at the end of Figure 10.16 indicate these further possibilities.

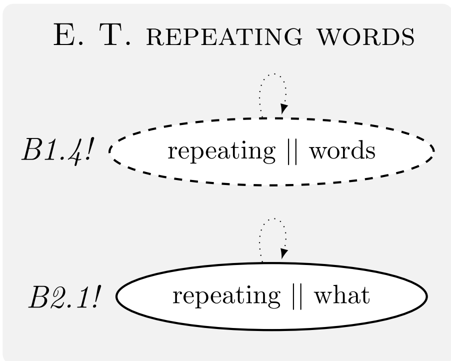
10.4.2.2 Variation derived from addition

Addition within variant strings

Addition of linguistic elements in responses is modelled by generating recognition paths that allow for the presence of unexpected elements. Figure 10.17 shows a recognition path that allows for the insertion of unexpected linguistic objects between the expected words – a recognition path that applies at the word level. The number



(a)



(b)

Figure 10.15: Expansion of RSL-based response patterns using omission as a transformation operation.

of loops that a recognition path can perform in one of the unexpected nodes can be determined.

In addition to the number of unexpected items allowed in between the expected words, the linguistic nature of the elements can also be restricted. For instance, in article-noun sequences it can be useful to allow for the presence of words with POS tags that typically appear in a noun phrase: adjectives or adjective-like words, adjective modifiers, and other sort of determiners. However, it is less motivated to include finite verbs, subordinating or coordinating conjunctions, simply because then we are less certain about the kind of linguistic structure recognised.

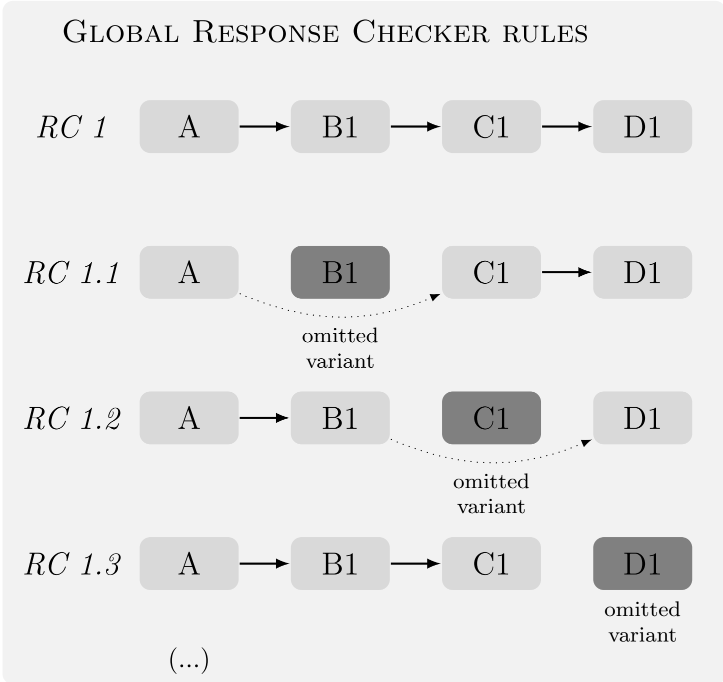


Figure 10.16: Recognition paths expanding one of the RC sequences of the E.T. comprehension activity.

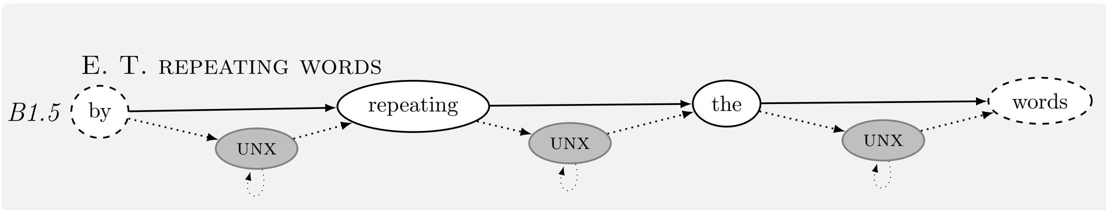


Figure 10.17: Paths generated by expanding B1 variant strings by means of addition operations.

With this strategy the path in Figure 10.17 could be converted to a path as the one reflected in Figure 10.18, where only certain determiners, those that do not have *the* as lemma information, adverbs and determiners are allowed between the tokens *the* and *words*.

Addition within response component sequences

To explain how addition as a transformation operation can be used to expand the range of possible RC sequences, we first present two sentences in (63), which deviate from the specified responses – unexpected contents in italics.

(63) a. By repeating the words it heard watching Sesame Street *in the living room.*

b. By repeating the words it heard watching Sesame Street *while drinking a cup of peppermint tea.*

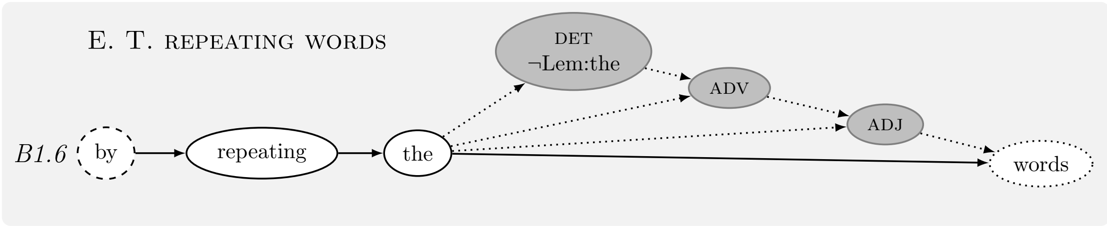
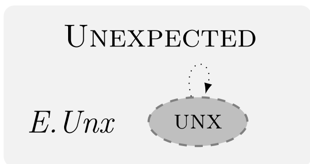


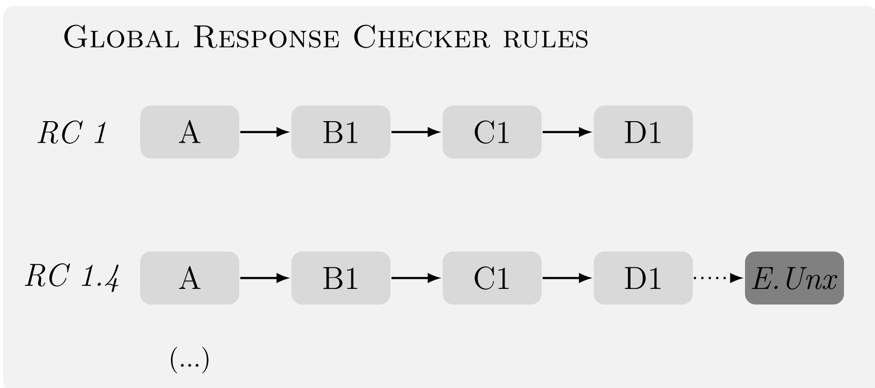
Figure 10.18: Paths generated by expanding B1 variant strings by means of POS-filtered addition operations.

The difference between (63a) and (63b) is that the former could be correct, even if not foreseen in the ReSS, while the latter would never be. The strategy we present allows for the detection of both types of deviations, but it cannot distinguish between semantically correct or situationally coherent deviations, and those that are somehow incorrect or incoherent.

These two deviating sentences require two types of modifications in the pattern recognition strategy: one at the IE modules level and one at the GRC level. To analyse chunks of unexpected words, a variant-level recognition pattern UNEXPECTED is needed: This is reflected in Figure 10.19a. Such a recognition path consists of only one node consuming any word that could not be taken by the other recognition paths.



(a)



(b)

Figure 10.19: Paths generated by expanding variant strings and RC sequences by means addition of unexpected linguistic items.

Figure 10.19b exemplifies the changes at the level of the Global Response Checker: This sample rule detects the combination of the expected response components and additional blocks with unexpected words. Given that unexpected elements can be found in any position in the response, several recognition paths at the level of GRC could be generated.

10.4.2.3 Variation derived from substitution

Substitution within variant strings

On the basis of POS-tagged text, the handling of substitutions can be realised at different levels. One of them is the replacing of one word, its word form, with a different word form with the same root, the same lemma. This is exemplified in the sentence fragments in (64) showing deviations in the string included in Variant B1 of the RC E.T. REPEATS WORDS. In (64a) and (64b) the expected form of the root *repeat*, *repeating*, has been replaced by forms of the same verb in other tenses; in (64c) the expected form of *word*, *word*, is presented in singular, not in plural.

(64) a. by *repeat* the words
  b. by *repeats* the words
  c. by repeating the *word*

The type of variation exemplified in (64) can be handled with recognition paths as the ones in Figure 10.20. This path allows for the analysis of responses where the expected word form of nouns and verbs in the variant has been replaced by another form with the same lemma and the same POS.

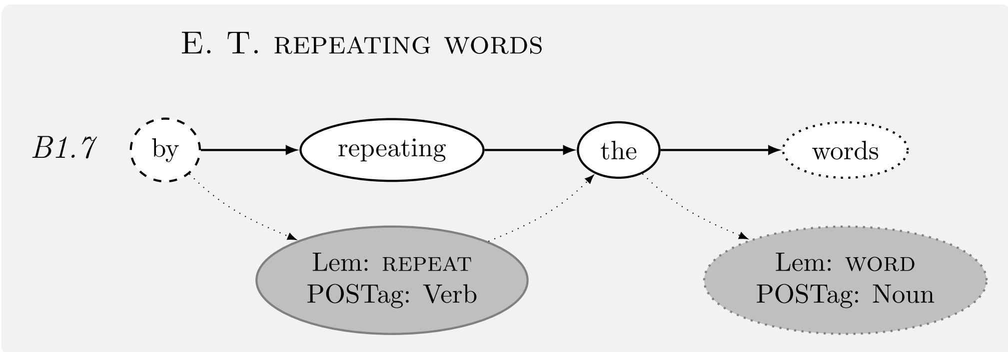


Figure 10.20: Paths generated by expanding B1 variant strings by means of addition operations.

Another type of substitution is the replacing of one word by a word with the same POS but with a different lemma, exemplified in the sentence fragments in (65) – where a function word is replaced with one that has the same POS but a different lemma. In (65a) and (65b), the preposition *by* is replaced with *in* and *through* respectively. In (65c) the determiner *the* is replaced with *these*.

(65) a. *in* repeating the words
  b. *through* repeating the words
  c. by repeats *these* words

The sentence fragments in (66) exemplify the replacing of one content word, as opposed to a function word, with one that has the same POS but a different lemma.

In (66a) and (66b), the verb *repeating* is replaced with *duplicating* and *putting* respectively. Both of them would result in incorrect responses, but the fragment using *duplicating* has a special relation to the original, since *repeat* and *duplicate* can be used as synonyms in certain contexts. In (66c) the noun *words* is replaced with *worlds*. Though it looks more as a typo than any other type of error, it results in a lemma-based substitution.

(66) a. by *duplicating* the words
b. by *putting* the words
c. by repeating the *worlds*

Figure 10.21 presents two recognition paths, B1.7 and B1.8, that would be generated on the basis of a linguistically analysed version of the string in B1 in Figure 10.12 to handle the variations represented in (65) and (66). The strategy presented is based on the recognition of nodes whose lemma is different from the expected one, but the POS tag is the expected one.

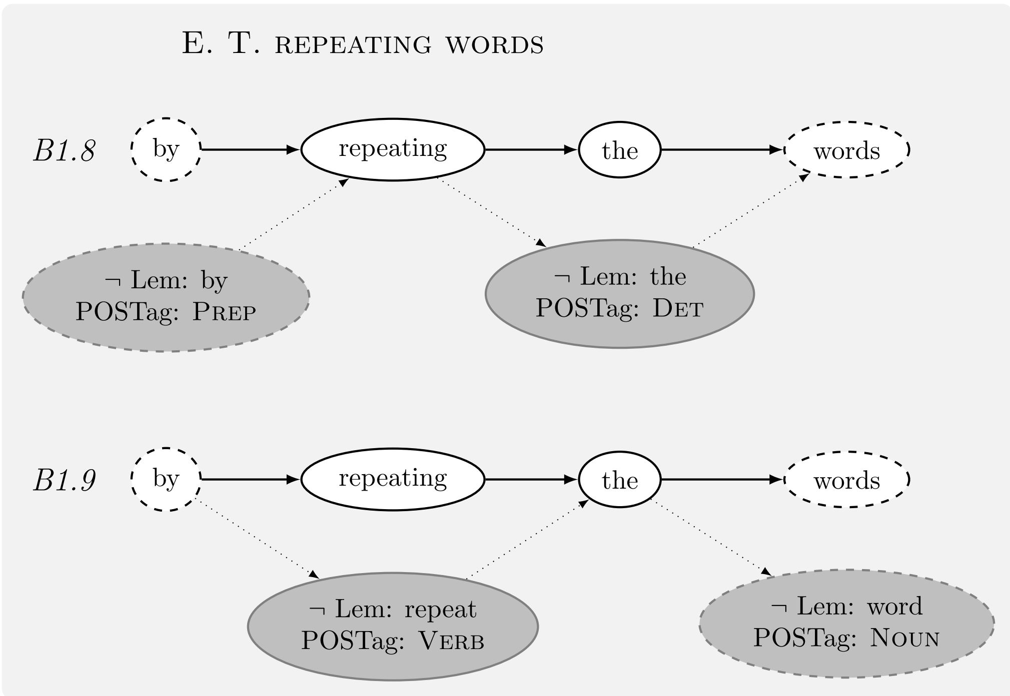


Figure 10.21: Paths generated by expanding B1 variant strings by means of substitution operations.

Substitution within response component sequences

Substitution of variants in RCs is handled as blending (see Section 10.4.2.5), since it supposes the use of two different realisations of the same concept presenting different linguistic forms. As such, these variants are not compatible in terms of syntagmatic relations with variants in other RCs.

10.4.2.4 Variation derived from reordering

Response variation derived from reordering can be found both at the level of variant and at the level of response component sequences. The sentence in (67a) exemplifies a deviation originated by a re-ordering within a variant: The order of the complementary clause *that it heard* is written as *that heard it.* The sentence in (67b) exemplifies a variation resulting from re-ordering at the level of response component sequence. Re-ordering does not necessarily imply an error. While (67a) can reasonably be marked as erroneous, at least formally speaking, (67b) is acceptable both in terms of thematic contents and in terms of linguistic contents.

(67) a. By repeating the words that *heard it* watching Sesame Street.
b. *Watching Sesame Street* by repeating the words it heard.

Reordering within variant strings

At the level of IE modules, this type of variation is handled with recognition paths that take into account certain pieces of linguistic information that allow for motivated order alternatives. The sentence in (67a) exemplifies a predicate-subject order change with respect to the expected response. Whether this is relevant for a particular target learner should be decided by the content designer, or even by the teacher – or using SLA references, and particularly learner corpus studies. However, this type of work falls out of the scope of this thesis.

Figure 10.22 would be generated by anticipating that subject and predicate can be presented in a different order. On the basis of POS tagged text, this is done using heuristics that check for the position of nouns and verbs within a sentence or sentence fragment.

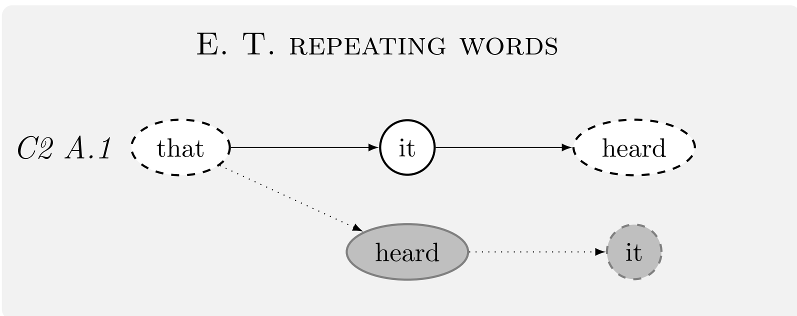


Figure 10.22: Paths generated by expanding C2 variant strings by means of reordering operations.

Reordering within response component sequences

To model responses as the one exemplified in (67b), new recognition paths have to be added in the Global Response Checker. For instance, Figure 10.23 shows a recognition pattern (RCS 2.1) that would process a response component order that is a variation of the RCS B1 → C1 → D2 (see Figure 10.12) and becomes D2 → B1 → C1.

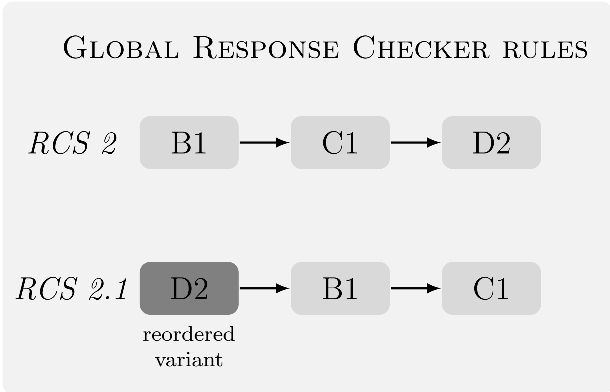


Figure 10.23: Recognition paths expanding one of the RC Sequences of the response to the E.T. comprehension activity.

10.4.2.5 Variation derived from blends

Variation based on blending can take place both at the level of variant and at the level of response component sequence.

Blends within variant strings

The sentence fragment in (68) reflects the use of the structure *heard to*, a blend between the structure *listen to* + ARG and *hear* + ARG.

(68) By repeating the words that it *heard to* watching Sesame Street in the living room.

Figure reflects 10.24 the recognition path to handle this deviating structure. It looks for a preposition *to* after the node containing *heard*.

Blends within response component sequences

The sentence in (69), already seen in (61), reflects a blend resulting from combining two response components that do no match. If the learner response has no reference to Elliot's sister in the part of the sentence previous to the text chunk *while she watched Sesame Street*, then this is not pragmatically acceptable.

(69) #*By repeating the words it heard while she watched Sesame Street.*

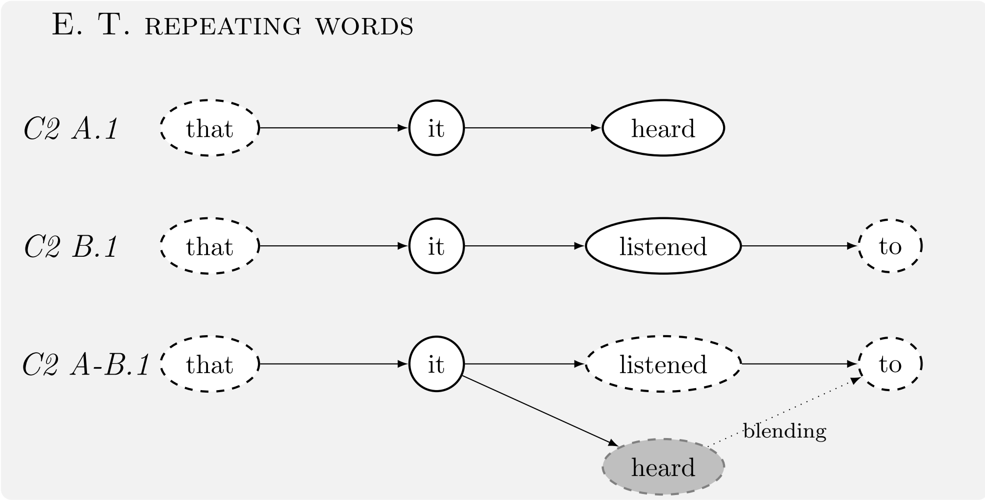


Figure 10.24: Paths generated by expanding C2 variant strings by means of blending as a transformation operation.

Figure 10.25 shows three recognition paths: The first two, RCS 1 and RCS 2, derive from the specified RCS (see Figure 10.12). The third one, RCS 1-2.1, responds to the use of variant D3, instead of D1 or D2, to express the concept related to RC D, SOMEONE WATCHED SESAME STREET. This combination is not possible because of the co-reference problem between *it* and *she*. The blending structure results from using a variant different from the expected as a continuation of the sequence A1? → B1, which can be followed by either D1 or D2, but not by D3.

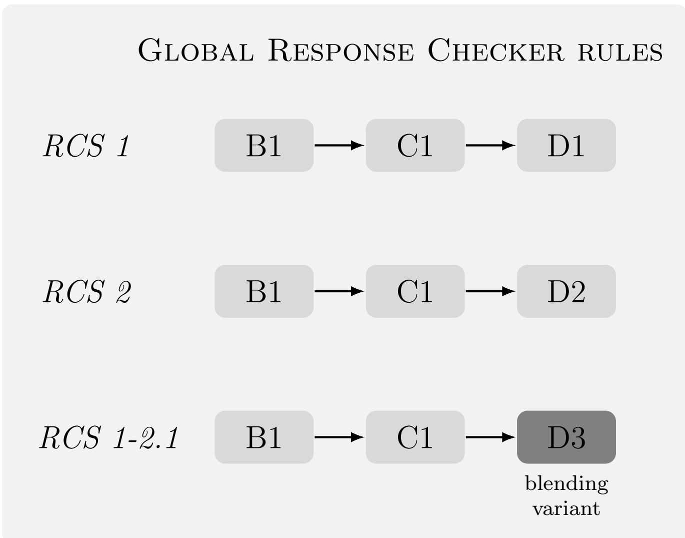


Figure 10.25: Recognition paths expanding one of the RC sequences from Figure 10.14 by allowing blending structures.

10.5 Chapter summary

In this chapter, we proposed the adaptation of the NLP-based feedback generation architecture described in Chapter 8 so that the domain-specific, that is, task-specific, modules in it can be designed and customised by FLTL content developers following specific design and response anticipation methodologies.

The customisation of the NLP-based feedback generation architecture requires the introduction of a pedagogical-computational interface. This interface is the Response Specification Language, which, captures the pedagogical and linguistic characteristics implicit in a list of correct responses to a given ICALL task and allows for the automatic generation of finite-state rules to be inserted in the customisable modules. The generation of the finite-state rules is based on the exploitation of two types of information: linguistic information and surface-transformation operations extracted from the analysis of the explicit and implicit structure and nature of the correct responses.

We presented the Response Specification Scheme, a methodology for teachers to specify a given list of gold-standard responses in terms of the Response Specification Language. The ReSS does not require computational expertise. It relies on general cognitive abilities, an understanding of the notions of syntagmatic and paradigmatic relations. Finally, we illustrated how the ReSS would be applied to a given ICALL task, and what the resulting finite-state rules would look like.

Chapter 11

Integrating ICALL in secondary education environments

This chapter presents our research to evaluate the integration of an ICALL activity authoring tool in secondary education instruction settings.[1] To do so, we integrate the RSL-based NLP resource generation strategy in a more general technical solution for educational purposes that allows for the use of the ReSS as a response specification procedure: This solution, AutoTutor, was implemented in the AutoLearn project.

We start by describing an experiment setup: the characteristics of the instruction setting, the participants and the technology – the learning management system – in which the ICALL authoring tool is integrated. We present the procedure by which the participating teachers and learners were trained and monitored. In the experiment, teachers are expected to design and implement FL learning materials including ICALL tasks, and learners are expected to work with these materials.

Following the presentation of the procedure, we present the results: the materials created by teachers and the learners' learning experience. To analyse teacher materials, we will take into account their pedagogical characteristics, as well as the integration of ICALL materials in a more general setting. As for the authored ICALL materials, we will specifically comment on their RIF characterisation, and on the nature and the linguistic complexity of the resulting specifications. To analyse the learner experience, we will evaluate the performance of the feedback generation system, and the degree to which there was evidence of learner uptake for a subgroup of the participants.

We conclude with a discussion in which we take into account the perspective of the teacher and the learner, by commenting on data gathered through questionnaires and interviews. The discussion includes the perspective of the researcher by analysing three aspects: the material creation process and its product, the use of the materials in class, and the limits of the current implementation of the NLP-based feedback generation strategy.

[1] This work was carried out within the framework of the ICE^3 project, with the acronym standing for Integration of CALL in Early Education Environments – a project funded by the Lifelong Learning Programme 2007- 2013 of the Education, Audiovisual and Culture Executive Agency, grant agreement 2010-3833/001-001.

11.1 Experiment setup

The setup of the experiment is designed to maximise the reality of the conditions of use of the technology. In this respect, we chose to work with secondary schools in which a minimum of CALL materials, or CALL-based instruction, is already in use.

11.1.1 Characterisation of the instruction setting

In order for schools to qualify for the experiment, we required them to comply with the following instructional and technical conditions.

Instructional conditions include the course setting, the pedagogical approach, the integration of the produced materials in the course syllabus, the type of FL learning activity, and the type of feedback to be generated. AutoTutor and its accompanying methodology are developed to be integrated in:

- Instruction settings where a blended learning approach is followed
- Courses following a communicative approach to language learning
- The creation of individual learning activities to support a particular syllabus
- The creation of Task Type I activities, as defined in Section 7.2.2.1 for the reasons argued in Section 10.2
- The generation of formative feedback

As for the **technical conditions**, AutoTutor was integrated in an existing Learning Management System (LMS), and made use of already available NLP tools. Thus, technically speaking is designed to:

- Be compatible with Moodle, a very popular[2] open-source web-based LMS that provides many functionalities and extensions useful both for foreign language teaching and for course management.
- Rely on the use of pattern recognition techniques such as finite-state automata, and exploit the power of mal-rule approaches to the analysis of learner language.

Both of these decisions relate to the necessity of developing a multi-platform, web-based, modular and robust software solution including state-of-the-art NLP technologies, one that is suitable for real-world instruction settings.

[2]According to http://moodle.org/ in March 2012 more than 67,000 registered Moodle sites existed around the world.

11.1.1.1 Expected user actions and roles

AutoTutor is expected to have two main uses: first, to enable FLTL practitioners to develop ICALL materials, and, second, to provide both teachers and learners with appropriate interfaces to use the teacher-developed ICALL materials in a blended learning context. We foresee see two possible user roles for teachers: the content developer and the course instructor. The learner user role is the material consumer.

For each of these three user roles we expect the following actions to be performed:

1. As a **content developer**, an AutoTutor user is able to:
   (a) Create, modify and remove ICALL learning activities
   (b) Specify the responses using the ReSS to generate the NLP resources for the task-specific response evaluation strategy for each activity item
   (c) Customise feedback messages for response patterns that can be anticipated as problematic for the target learner
2. As a **course instructor**, an AutoTutor user is able to:
   (a) Upload an ICALL activity as a Moodle activity[3]
   (b) Monitor learner activity
   (c) Monitor performance of assessment functionalities
   (d) Comment on the system's feedback and on learner responses
3. As a **learner**, an AutoTutor user is able to:
   (a) Perform learning activities
   (b) Obtain immediate NLP-based automated formative feedback
   (c) Keep track of his or her own previous activity

These actions correspond with two basic software components: an **activity editor** and an **activity player**.[4] The natural user of the activity editor is the content developer (an FLTL practitioner, a teacher), while the natural users of the activity player are the instruction teacher and the learner. In our setting, content creator and instruction teacher will be the same individual. Though each of the corresponding roles is performed through different components of AutoTutor, or through different functionalities of Moodle, this thesis will concentrate on the characteristics and the use of the software components that facilitate the authoring of ICALL materials.

[3]In Moodle the term *activity* is used to refer to those materials where learners have to "do" something, while the term *resource* is used to refer to those materials considered contents that are provided to learners as input requiring no further perceptible action.

[4]The editor/player dichotomy is a common concept in software solutions that require a component to author documents and a component to employ them.

11.1.2 Participants

Three secondary school teachers participated in this experiment. Two of the teachers work in the school Fundació GEM in Mataró, Catalunya, Spain; the third teacher works in the school Fundació Llor in Sant Boi de Llobregat, located also in Catalunya. Both schools are *escoles concertades*, that is, they are partially funded by the Catalan government, which makes them neither fully public nor fully private.

The teacher from the school in Sant Boi de Llobregat teaches *Science*, therefore, his teaching can be considered Content and Language Integrated Learning (CLIL). The two teachers from the school in Mataró are FLTL teachers, and both teach English as a Second Language. These two teachers decided to work in collaboration for the purpose of the experiment.

The three teachers teach learners in Educació Secundària Obligatòria (Compulsory Secondary Education). The two teachers in Mataró teach also in Primer de Batxillerat (literally, First of Baccalaureate), the first year of non-compulsory secondary education, which learners opting to go to the university need to complete.

11.1.2.1 Teacher profiles

The teacher from the school in Sant Boi de Llobregat, referred to from now on as Teacher 1 (T1), is a male in his mid-40s who has been teaching for more than 15 years. He has always taught subjects related to science and technology both in primary and secondary education. As a computer user, T1 has a highly qualified profile: He uses computers both for personal and professional activities daily, including the use of communication software (e-mail, chats), search engines, multimedia tools, web 2.0 tools, and productivity software. Moreover, T1 has a proactive attitude toward integration and experimentation with new methodologies and technologies. This has pushed him, for instance, to be one of the first teachers in his school to teach CLIL (starting in 2009), as well as to pioneer the use of Moodle as a platform to regularly teach and communicate with learners.

The two teachers from the school in Mataró, referred to from now on as Teachers 2 and 3 (T2 and T3), are female teachers; T2 is in her early 40's, and T3 in her mid-40's. T2 and T3 have always taught English as a Second Language, while T3 has also taught German as a foreign language to beginner learners. Both of them have taught for more than 15 years in primary and secondary education. As computer users, T2 and T3 have a highly qualified profile: Both of them use computers both for personal and professional activities daily, including the use of communication software (e-mail, chats, micro-blogs), search engines, multimedia tools, web 2.0 tools, and productivity software. Moreover, both T2 and T3 have a proactive attitude toward integration and experimentation with new methodologies and technologies. For instance, T2 and T3 have introduced the use of Moodle as an LMS in their school, becoming de facto Moodle administrators.

11.1.2.1.1 Setting and approach of Teacher 1

T1 describes his instruction context as one in which learners present quite different levels of proficiency in their foreign language, which he interprets as a need to foresee strategies and activities that allow him and his learners to follow different learning paces and paths. T1 says he uses both Catalan and English during explanations, with the justification that certain learners would not be able to follow the explanation in English. However, group and individual activities are exclusively in English. In terms of infrastructure, some of the learners have their own computer, but not all of them. The learners who participated in the experiment went to the school's PC lab for some of the activities.

T1 describes his approach as based on competencies, where topics are more relevant than textbook lessons. His syllabus is based on the programme established by the regional Ministry of Education, but he actually uses a blend of available materials and materials he created. As for published materials, he includes in this category open materials or real-life materials (news, handbooks, and so on). He shares experiences with other teachers in his school and seeks to integrate positive experiences, particularly in terms of group activities. He collaborates with teachers from other subjects to create cross-disciplinary and interconnected materials. For instance, in the case of the materials for the subject *Science* that he develops for this experiment, he coordinates the linguistic contents of the units of work with teachers from the English Department.

In terms of supplementary materials, T1 states that he organises his classes so that most of the work is done in class, and learners are encouraged to do so. However, if learners cannot finish their work in class, they do it as homework. T1 states that approximately 50% of the homework that his learners are given is delivered and accomplished by means of a computer.

When asked about how he perceives the way learners use the feedback he provides, he says, "They usually don't read it; most of the time they only care about finishing the activity as soon as possible". According to him, for many of his learners, "Every new thinking task is a new boring task".

11.1.2.1.2 Setting and approach of Teachers 2 and 3

T2 and T3 describe their instruction context as one in which learners present different levels of proficiency in their foreign language. They overcome this situation by grouping learners by proficiency levels and using different criteria with learners with different levels. Both teachers use only English in class. In terms of infrastructure, none of the learners has his or her own computer, so they all need to go to the PC lab for CALL activities or they do them at home.

T2 and T3 describe their approach to language teaching as one where all skills are worked on in the same degree, as well as one that encourages learners to put into practice everything they know about the foreign language. They try to organise dynamic units of work and relate the contents of the activities with the learners' own experiences. With best performing learners they prepare activities that learners can use in other areas by using text from newspaper, cultural events or fiction. Their syllabus is based on the Ministry's requirements, though they try to adapt it to the group needs, and tend to skip contents they may already know. They enhance ministry-based curriculum with extra materials created by them or third parties, and include project-based learning activities in the course.

In terms of supplementary materials, T2 and T3 state they give learners a bit of homework every week, of which roughly between 25% and 50% is computer based. T2 and T3 state that writing is the skill that "you always have the feeling you could have practised more". Moreover, they complain about not having enough time to practice and evaluate writing tasks as regularly.

When asked about how they perceive the way learners use the feedback they state that most learners seem not care about it, though they really cannot tell whether their feedback is useful or useless for the learners. When they give feedback to the learners, they devote part of the class to encouraging learners to have a look at it and ask questions: "Hopefully they take it into account in their future linguistic performances."

11.1.2.2 Learner profiles

11.1.2.2.1 Groups of T1

Teacher 1 involved three learner groups in the third year of obligatory secondary education, which we will call 3A, 3B, and 3C. As for the number of learners, 3A and 3B have 27 learners each, and 3C has 24. The distribution of male and female students is quite balanced: 40% female and 60% male students for 3A, and 55% female and 45% male students for groups 3B and 3C. Learner ages are between 14 and 15 for the three groups, except for one student in 3A who is 16. Their mother tongues are Catalan and Spanish: A percentage of learners feels more comfortable with Catalan than with Spanish (60%), and another percentage feels more comfortable with Spanish (40%). Three of them spoke other languages at home (English, German and Romanian). 83% of the learners had been learning English for 7 or more years at school, 12% had been learning English for a period of 4 to 6 years, and the other 5% had been learning it between 1 and 3 years.

As computer users, the learners in groups 3A, 3B and 3C are frequent users of communication and entertainment applications: 80% of them use the computer to communicate with others on a daily basis, and 68% state they use it for entertainment purposes. In contrast, 21% state that they use the computer on a daily basis at school, and 29% state that they use it at home for learning.

11.1.2.2.2 Groups of T2 and T3

Teachers 2 and 3 involved four learner groups. Two of them are groups of learners in their first year of obligatory secondary education, which we will call 1A and 1B. The other two are groups of learners in their second year of obligatory secondary education, which we will call 2A and 2B. Groups with the suffix A, (1A, 2A) worked with T3, and groups with the suffix B worked with T2.

Group 1A has 20 learners, group 1B and 2A have 22 learners each, and group 2B has 12. The distribution of male and female students is quite balanced: 36% female and 64% male students for 1A, 55% female and 45% male students for group 1B, 55% and 45% for group 2A and 58% and 42% for group 2B. In groups 1A and 1B learner ages are between 12 and 13, and in groups 2A and 2B learner ages are between 13 and 14. Their mother tongues are Catalan and Spanish, though the percentage of learners more comfortable with Catalan is higher than the percentage of learner more comfortable with Spanish (80% vs. 20%). Two of the learners spoke amazique or/and berber at home. For groups 1A and 1B, 44% of the learners had been learning English for 7 or more years at school, 50% had been learning English for a period of 4 to 6 years, and the other 6% had been learning it between 1 and 3 years. For groups 2A and 2B, 71% of the learners had been learning English for 7 or more years at school, 24% had been learning English for a period of 4 to 6 years, and the other 5% had been learning it between 1 and 3 years.

As computer users, 64% of the learners in groups 2A and 2B use computers for communication purposes, and 58% of them use them for entertainment. 23% of the learners state that they use computers for learning, while only 17% of them use them almost daily at school.

11.1.3 An authoring tool for ICALL activities

As a software solution AutoTutor is more than an ICALL authoring tool, it is a toolkit to author and manage ICALL activities within Moodle. Our interest, and our contribution, is that we enhanced AutoTutor with a component that allows for the generation of teacher-driven NLP resources that are later used in the automatic feedback generation functionality. This functionality and the response specification and NLP resource generation process is based on the customisable NLP-based feedback generation strategy described in Chapter 10.

As already mentioned, AutoTutor consists of an activity player and an activity editor, and we will focus here on the latter. The activity editor is the part of the technology that empowers teachers to author ICALL activities without the intervention of a computational linguist. Along our explanations reference and background information will be provided on the activity player if it is required to better understand the characteristics of the experiment.

AutoTutor's activity editor is called AutoTutor's Activity Creation Kit (ATACK). ATACK is a piece of software for content designers to create HTML pages – with text, image, video and audio –, including activity instructions and a related set of questions. For each of the questions a set of correct responses can be specified following the ReSS. ATACK guides content creators through the process of applying the ReSS. Afterwards, per button-click, it automatically expands the provided specifications to generate the NLP resources for the assessment of learner responses.

11.1.3.1 Graphical User Interface

ATACK's Graphical User Interface (GUI) consists of two areas: an area where global activity characteristics can be defined, and an area where questions for the activity are inserted – including feedback specifications. ATACK's interface is not developed by design specialists. It just contains the functionalities needed to proof the concept that teachers can author and manage ICALL materials with the ReSS and the accompanying technology.

11.1.3.1.1 Global activity actions in ATACK's GUI

Figure 11.1 shows the area for the definition of the global characteristics of activities. This includes the possibility to manage AutoTutor files, like in the *File* menu option of most editors, and to perform actions related to the usage of the activities.

As for file management options, content authors can create a new activity through the button *Create exercise*, cancel this action or closing the activity through the *Cancel* button, open an existing activity through the *Load exercise* button and the accompanying drop-down list. Users can also save activities using the *Save* button and give them a name in the *Exercise name* text area.

As for the activity specific actions, users can specify the foreign language for which the activity will be produced, which affects the way the NLP resources will be automatically generated, but also the language in which the interface and the feedback messages are shown to learners. In its current version AutoTutor activities always use the corresponding foreign language in these two respects.[5]

As part of the activity usage actions, users can prepare the activity to be uploaded in the corresponding Moodle server: In this case by saving it in HTML format using the *Generate HTML* button. And, finally, users can generate the NLP resources required for enhancing the HTML pages with NLP-based feedback generation functionalities using the *Generate grammar* button.

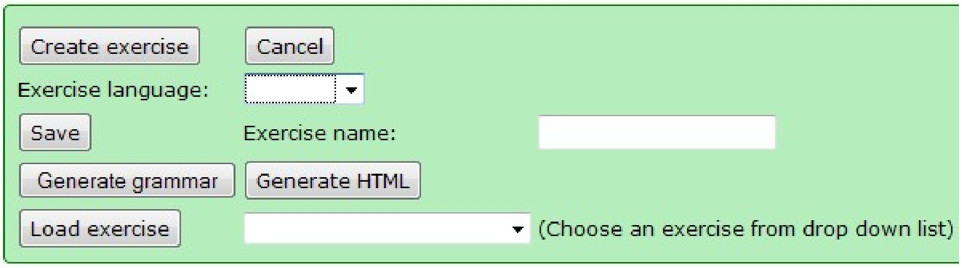


Figure 11.1: ATACK's GUI: settings and global activity actions

11.1.3.1.2 Question insertion and response specification

Figure 11.2 shows the first view of the graphical interface accessed by content designers for the insertion of activity items (questions) and response specifications. Content designers can use the *Add question* and *Delete question* buttons to add or delete a question. Question 1 in Figure 11.2 shows a text, a question, that will be shown to the learner for him or her to produce a response. The tab *Question 1* has response and feedback specification areas: *Create blocks*, *Organise blocks*, *Customise feedback* and *Sample answers*.

From left to right, the first two tabs in Figure 11.2, *Create blocks* and *Organise blocks* are the ones where content developers detail Response Components, through specifying variants and strings in them, and Response Component Sequences.

[5] AutoTutor allows for the generation of activities in English, German and Spanish, but for the purposes of this thesis only activities where English is a foreign language will be considered.

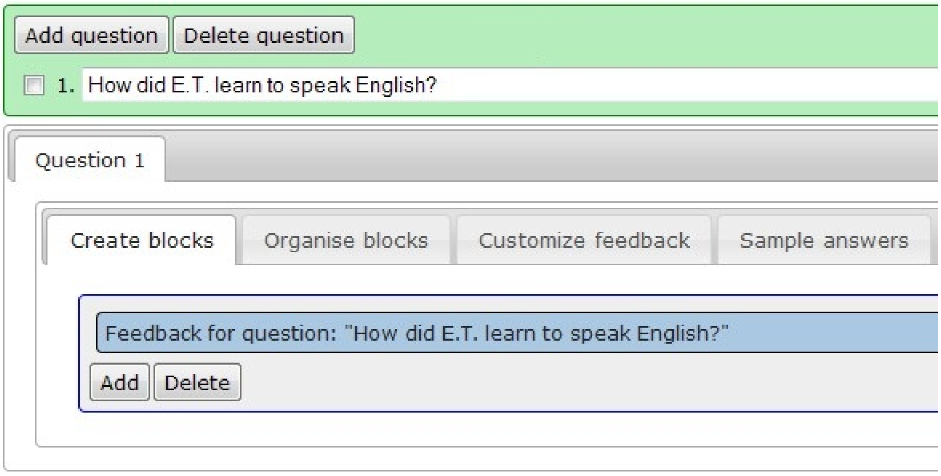


Figure 11.2: ATACK's GUI: activity authoring area

The *Customise feedback* tab allows teachers to determine specific linguistic expressions for which a specific feedback message should be generated, and for which the automatically generated feedback is not satisfactory to them. Finally, the *Sample answers* tab is for teachers to provide AutoTutor with a list of sample correct responses to the question that can be visualised by the learner after the system has shown the correction feedback.

Introducing Response Components Figure 11.3 shows how the results of introducing the elements of the Response Component E.T. LEARNT ENGLISH for our E.T.-learns-English example in the previous chapter. Response Components are called *Blocks* in the GUI. Each of them has an associated text area where the title or a brief notional description can be introduced.

The blue rectangle with rounded corners contains all the information at the level of RC. The *Type* option allows for the specification of two kinds of RC. Plain blocks contain lists of strings to handle text chunks that correspond to variants. This is the default and the normal characteristic for blocks. List blocks create variants that are enumerations. It requires the number of items expected in the enumeration, as well as the specification of the words that can be used to express each of the items in the enumeration.

The yellow area in Figure 11.3 corresponds to Variants, each of which consists of sets of Strings, in this case exemplified by the elements of the RC E.T. LEARNT ENGLISH. This area includes an *Add variant* and a *Delete variant* button. The deactivated *New item* and *Delete item* buttons, as well as the deactivated *Min* and *Max* fields, are only relevant for response components of the type list. Through them one determines the number of items in an enumeration as well as the different ways of expressing each item.

Preceded by the label *Text*, the yellow area in Figure 11.3 contains the list of text chunks associated with the only variant of the corresponding response component. Note that optionality is marked by explicitly introducing each text chunk versions in a separate *Text* field, and not using the slash as we did in the ReSS.

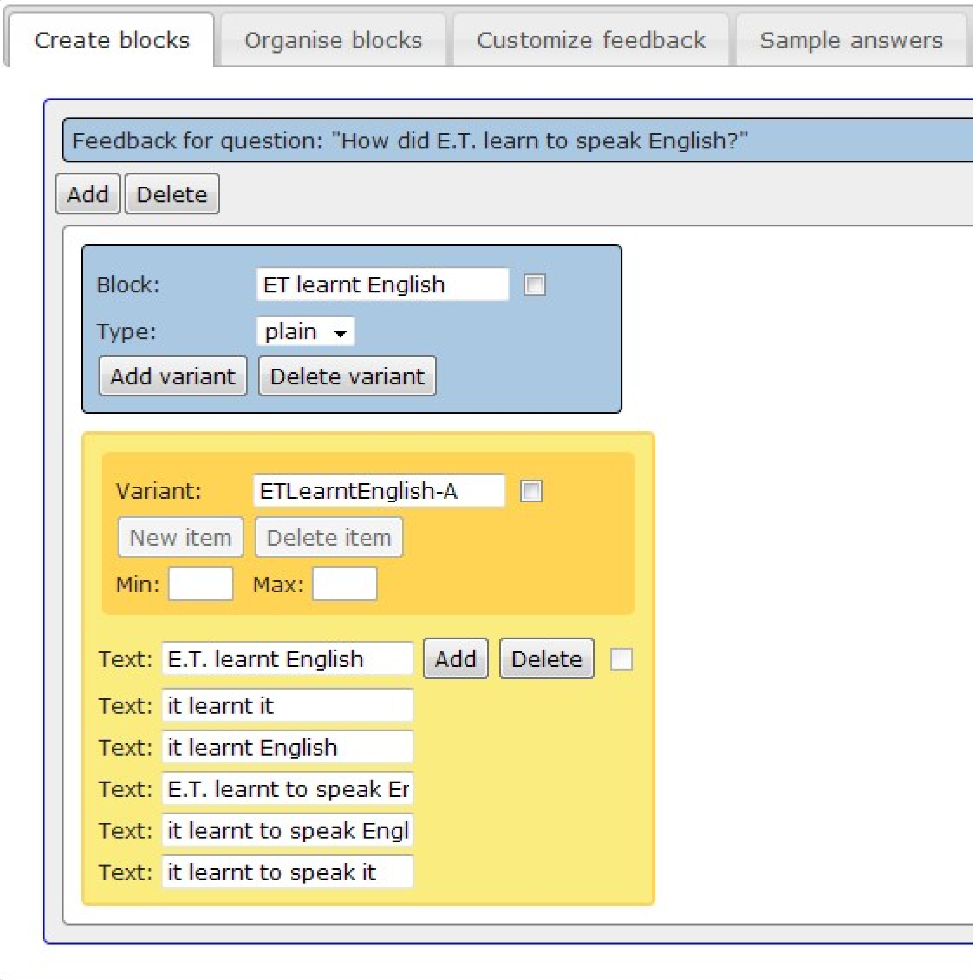


Figure 11.3: ATACK's GUI: question tab area to specify information for response components.

Figure 11.4 shows the information of the response component E.T. REPEATS WORDS once inserted in the interface. This RC has two variants, as we explained in Section 10.3.2.

Introducing Response Components Sequences The *Order blocks* tab in Figure 11.5 reflects the introduction of RCSs. The light grey area contains all the RC and the corresponding Variants – still using our E.T. activity example. The two light olive green areas contain specific RCS: A – B1 – C3 – D2 and A – B2 – C1 [...] – the second is unfinished. By dragging RC boxes from the light grey area and dropping them in the light olive green areas, content designers determine valid RCS. The *Add* and *Delete* buttons can be used to add or delete a sequence.

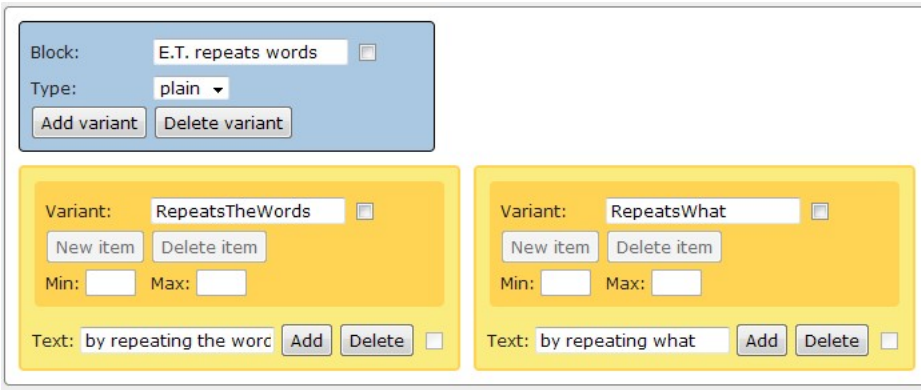


Figure 11.4: ATACK's GUI: specification of the information for response component E.T. REPEATS WORDS.

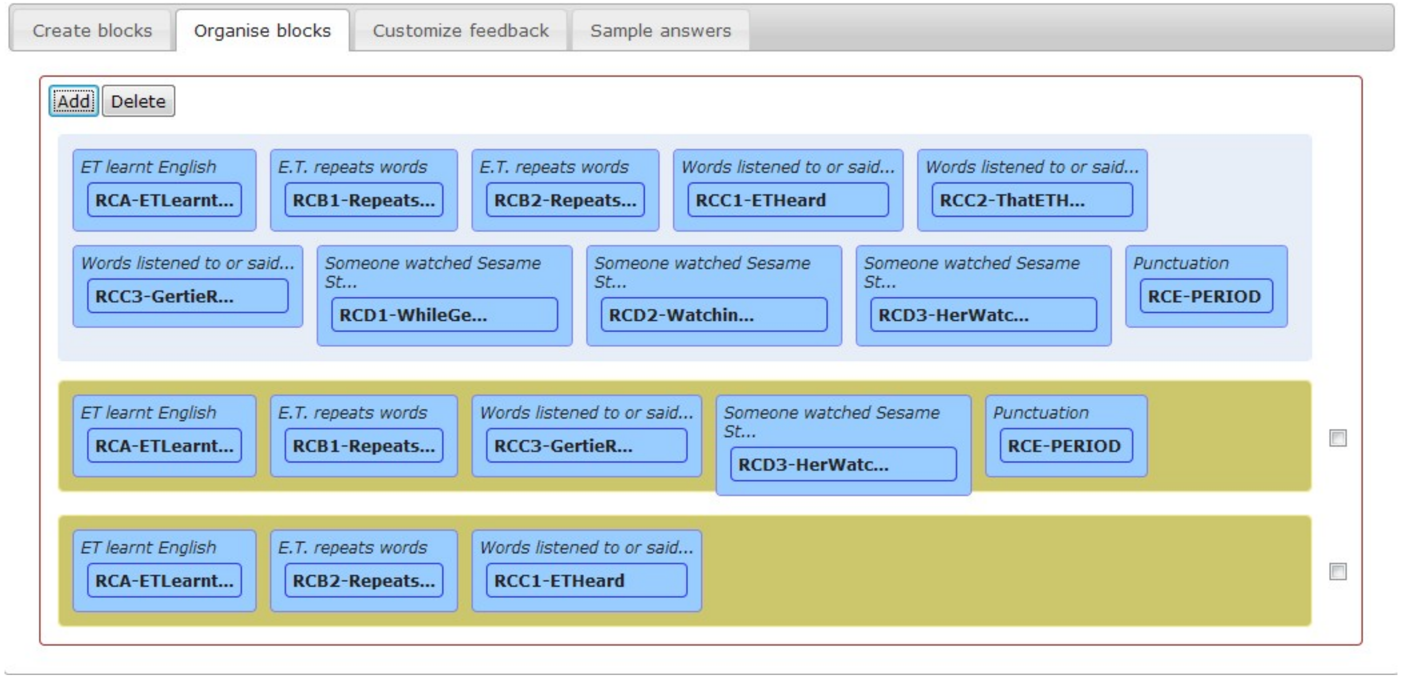


Figure 11.5: ATACK's GUI: ordering of the response components to build the sequences yielding correct responses.

11.1.3.1.3 An area for the insertion of specific feedback messages

Content designers can generate specific feedback messages for particular linguistic constructions using the *Customised feedback* tab. As shown in Figure 11.6, the information provided is a *Trigger* expression, that is, an expression that triggers the feedback message, and a *Message* to be shown.

The *Exact matching* option determines if the feedback message has to be shown only if the trigger expression is literally found in the learner response, or, if it can be shown also when the contents in the *Trigger* field are resolved at the lemma level.

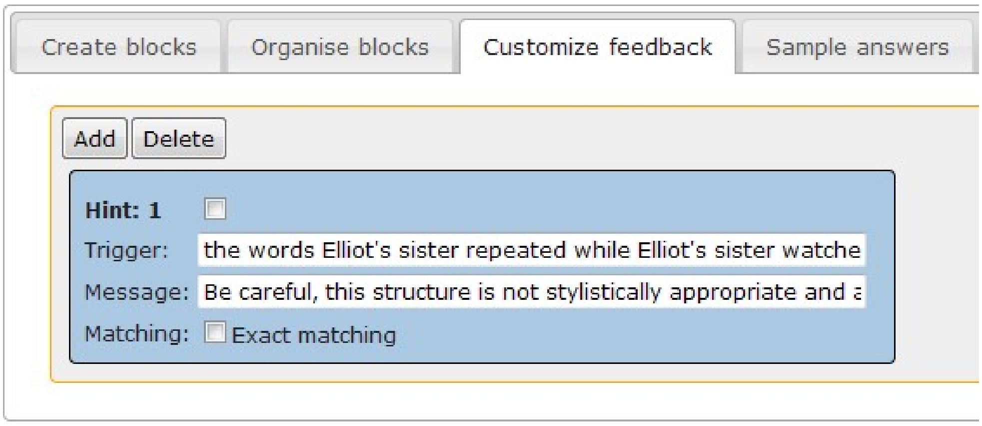


Figure 11.6: ATACK's GUI: *Customised feedback* tab, an area to define learner-oriented feedback messages.

This functionality is designed for teachers to be able to enhance system feedback. Through it they can provide learners with specific feedback messages associated with certain linguistic patterns found in the learner responses, patterns they might have observed and were not satisfactorily handled by the system.

11.1.3.1.4 An area for the insertion of sample answers

As shown in Figure 11.7, the *Sample answers* tab allows content designers to insert a list of sample answers, so that learners can have access to an approved solution. The list of sample answers is shown to the learners in case they do not respond correctly to the activity. This functionality is motivated by one of Chapelle's hypothesis regarding the musts for CALL materials (Chapelle, 1998). If teachers use it, the list of sample responses is converted into an HTML page that is separate from the ICALL activity page. If teachers upload it to Moodle, sample responses can be accessed by learners using the *Show me the answer* button in the activity player.

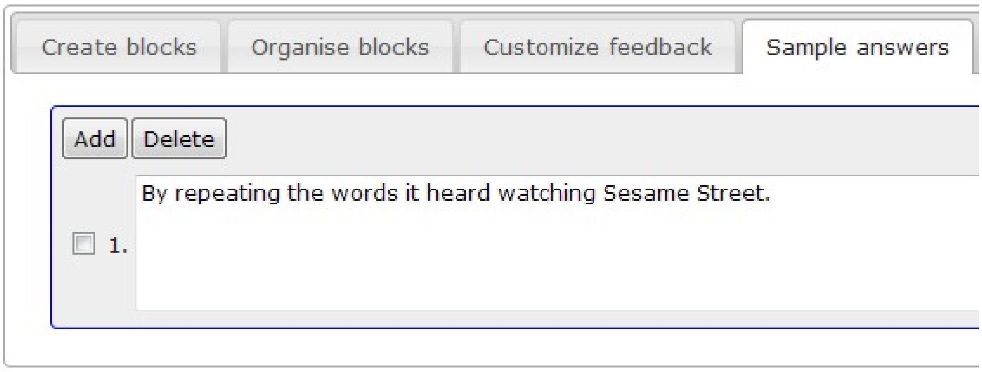


Figure 11.7: ATACK's GUI: *Sample answers* tab, an area to insert sample answers to be shown as part of the feedback.

11.1.3.2 Automatic generation of NLP resources

AutoTutor includes an implementation for the customisable automatic feedback generation strategy presented in the previous chapter, and on the NLP-based feedback generation architecture presented in Section 6.3.2.1 The solution is based on the processing formalisms MPRO and KURD.

Figure 11.8 shows the relationship between RSL-based response specifications and the actual implementation of the NLP resource generation strategy. The upper blue rectangles in the figure show the three different software components in which either response specifications or NLP resources are being defined, processed or used. The methodological and operational counterpart of each of these software components is described in the black boxes at the bottom of the figure.

The most-left component, ATACK's GUI is related to the methodological step that involves response specifications using the ReSS. As a result, RCs, RCSs, and teacher-defined error patterns can be used to generate NLP resources. This is part of the activity design and development process to be performed by the content designer. The middle component, the back-end of ATACK, is responsible for the generation of the customised NLP resources. This component relates to the technology that enables teachers to generate the NLP and response evaluation resources by customising a more general formative feedback generation strategy. The third component, the back-end of the activity player, uses the generated NLP resources for the actual evaluation of learner responses (in the figure ATAP stands for AutoTutor Activity Player).

The three processes in the middle of Figure 11.8 determine how specific elements of the response specification generate specific NLP resources for the customised assessment strategy. The back-end of ATACK maps the information specified by content designers and generates the domain-dependent NLP resources. Different elements of the ReSS are used to generate different kinds of NLP resources. While the specified response components are used to generate the pattern recognition paths of the Content Analyser, the response component sequences are used to generate the rules required for the checking of overall response correctness completeness in the Feedback Generator. Finally, the teacher-defined error patterns are converted

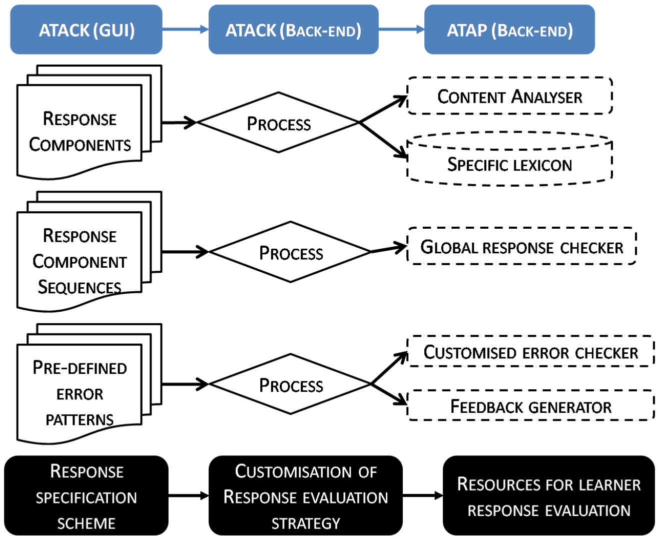


Figure 11.8: From the specifications provided by content designers in ATACK to the NLP resources needed by ATAP to evaluate learner responses.

into rules included in the Customised Error Checker. The Content Analyser and the Customised Error Checker are modules included later in the Information Extraction Modules of the customisable architecture, while the Feedback Generator becomes part of the Global Response Checker (cf. Figure 10.2).

The following three sections describe the technical details of the NLP resource generation process. A previous description of AutoTutor as an authoring tool was published in Quixal, Preuß, Boullosa, and García-Narbona (2010).

11.1.3.2.1 Generating the rules for the Content Analyser

The Content Analyser is expected to detect the specified response components using a series of form-based recognition patterns that include and expand the provided response specifications. As a linguistic analysis module, the Content Analyser will be responsible for the analysis of the task-specific thematic and linguistic contents.

Enriching teacher specifications with POS tagging Figure 11.9 reflects the process that undergoes each of the strings specified for each of the variants in a response component in order to be converted into a series of recognition patterns to check for exact or partial matches in learner responses. As shown in the first column in Figure 11.9, the process initially implies a standard POS tagging process, from which an annotated version of the text chunk is obtained. POS tagged strings contain information related to the word, its surface form, the lemma, the grammatical category, and morphosyntactic information such as mode, tense, number, gender, case, and so on. The kind of morphosyntactic information associated with each word is dependent on the grammatical category and is determined by the actual annotation tools (see Section 6.3.2).

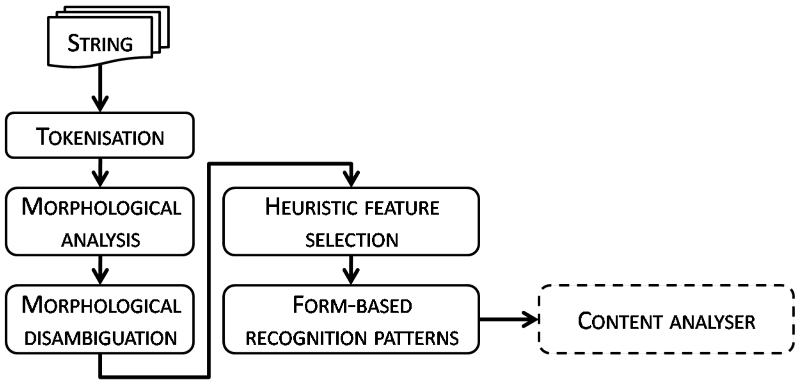


Figure 11.9: Software for the generation of expanded form-based recognition paths in the Content Analyser.

Heuristic selection based on linguistic properties The second step in the generation and expansion of the recognition patterns consists in determining a series of features found in the analysed string tokens that are known to be relevant for anticipating a series of variations to the linguistic structures underlying the specified variant strings.

The concrete expansion criteria implemented in AutoTutor are reflected in Table 11.1. The table has three columns: The first one identifies a series of linguistic structures for which surface transformation operations are implemented. The other three columns describe which of the possible transformation operations is applied.

As shown in column two, omissions are handled for many of the linguistic structures. For instance, the first two rows reflect that the RC strings containing definite and indefinite articles are expanded to detect responses where the infinite or definite articles are omitted. Similarly, the 10th row, related to composite verb forms, shows how the absence of inflectional morphology is detected – this includes third person singular inflection (-s), present and past participle (-ing and -ed), etc..

The third column shows the linguistic structures for which expansion based on addition is applied. Four of them foresee the insertion of unexpected elements between frequent or fixed syntactic structures marked with the label *In between*. For instance, the last three columns indicate that for *Determiner + Noun*, *Adjective + Noun*, and *Verb + Determiener/Adjective/Noun* sequences, the insertion of certain elements, particularly words expected within a noun phrase, is modelled.

The fourth column shows expansion techniques based on substitutions. This technique is the one in which linguistic criteria are applied to restrict the range of operations that can be generated. For instance, the first two rows alternate between the use of *a/an* and the use of *the*: This implies the detection of the correct uses of both articles and expect the use of the unexpected alternative. Other substitution based expansions are not lexically driven, but morphologically or syntactically. This is the case for the use of alternative number, as indicated in the third row for bare nouns, or the case of the use of base, comparative or superlative forms of adjectives, as indicated in the 13th row.

Admittedly, these feature selection heuristics would be more appropriately defined on the basis of FLTL or SLA studies. For the purpose of this research it is enough

Linguistic structure	Omission	Addition	Substitution
Thematic contents			
Unspecified noun, verb or adjective	–	UnxPlus	–
Specified noun, verb or adjective	∅	–	UnxOther
Linguistic contents			
Definite articles	∅	–	*a*, *an*
Indefinite articles	∅	–	*the*
Bare nouns	–	*the*, *a* or *an*	Noun in singular or plural
Possessive *'s*, *s'*	∅	–	*s'*, *'s*
Preposition	∅	–	Other prepositions
Modal verbs	∅	–	Other modals/verbs
Auxiliary verbs	∅	–	Other auxiliaries/verbs
Conjunction *that*	∅	–	–
Commas, *and*, *or*	∅	–	*and* vs. *or*
Inflected verb forms	No inflection	–	A different inflection
Composite forms: *will* + INF, *to* + INF etc.	–	In between	–
Personal pronouns	∅	–	*<I,me>*, *<he,him>*, ...
Adjectives	–	–	Base, comparative and superlative forms
Adverbs/Adjectives	–	–	Adjectival vs. adverbial form
Quantifiers	∅	–	Count/Mass distinction
Quantifiers	∅	–	Other quantifiers or determiners
Determiner + Noun	–	In between	–
Adjective + Noun	–	In between	–
Verb + Word	–	In between	–

Table 11.1: Heuristics for the selection of linguistic features in the reference texts for the expansion of the response components

to show that the linguistically motivated expansion of response specifications can predict variation patterns often seen in learner language.

11.2 Procedure

This section describes how teachers were trained and monitored to create and use ICALL activities in the respective instruction settings. After teachers agreed to collaborate, the experiment ran in three phases. The first one was the training of the participating teachers in the use of the methodology and the software to author ICALL activities. This training included methodological aspects related to task design, response specification, and some basic knowledge in Natural Language Processing. This phase expanded over approximately two months, in form of two- to four-hour sessions.

The second phase was the actual design and planning the use of ICALL activities, where teachers were expected to work autonomously, though they were technically assisted. Teachers were required to produce a set of materials integrated in their course programme including a minimum of three ICALL activities using AutoTutor and, optionally, another three CALL activities using Hot Potatoes. The overall work plan should foresee approximately eight hours of learner work.

The third phase was devoted to the use of the created materials with their learners. During this period teachers were given technical and practical assistance: Particularly in the initial session, we assisted them in the presentation of the experiment to the learners, as well as on the training of the learners of the use of the learner interface. This phase included the collection of profile and satisfaction questionnaires from learners, which were later used to evaluate the experience.

11.2.1 Teacher training

The training of teachers aimed to familiarise them with some basic concepts of Natural Language Processing, the response specification process, and the mechanics of using the interface to specify expected responses and to generate the corresponding NLP resources. The training included a review of key aspects of the design of FLTL materials and CALL materials, as well as some general aspects of content management in Moodle.

The training took place in the form of a one-day meeting where the morning session was devoted to introducing the nature of the collaboration proposal, as well as the general pedagogical and computational aspects of the work. The afternoon session was devoted to the mechanics of using AutoTutor as an authoring tool, as well as to the mechanics of managing AutoTutor activities in Moodle. In the following sections we detail the contents of this initial session.

11.2.1.1 Introduction to the experiment

This session introduced the project goals to the teachers. We emphasised that this experiment was aimed at the creation of CALL and particularly ICALL materials to be integrated in their learning programmes. Ideally, the created materials had to be completed by learners as supplementary work aiming to promote individual learning.

11.2.1.2 Pedagogical background and activity design

This session framed the pedagogical assumptions of the experiment and the basics of class material design. We argued for the need to carefully integrate ICALL activities in wider task-based instruction activities to keep in line with the communicative approach. Moreover we explained that ICALL activities offer learners the opportunity to obtain immediate feedback to their responses and to become aware of their errors and their learning needs. Teachers were offered the chance to work in collaboration and to seize the opportunity of being involved in this project to handle any question, doubt, or obstacle in the common monitoring sessions.

As for activity design, teachers we proposed that teachers:

Choose the learning unit Decide the learning unit in which the materials would be included, taking into account that the materials were going to be used around April-May 2011.

Identify the learning objectives Determine the goal of the activity in pedagogical terms.

Include CALL tasks in the pedagogical plan Decide when and how CALL activities would be incorporated in the lesson, and plan its pedagogical objective. That would include determining the types of exercises that were more appropriate for the learning objectives.

Use authentic CALL materials Search for authentic materials, and try to make them match with a communication context similar to the 'real-life' communication contexts.

Author CALL activity Identify the appropriate technique. Use either CALL (Hot Potatoes) or ICALL (AutoTutor) activities to implement it. For this, teachers were given a table in which activities could briefly be characterised and classified, as the one in Section G.1 in Appendix G.

Personalise learning content Adapt the contents to the needs of the group of learners by fine-tuning prompts, instructions, input data, scaffolding techniques in the input data, and so on.

11.2.1.3 Automatic feedback for assessment purposes

This presentation introduced teachers to the differences between the kind of automatic feedback generated by CALL activities with no linguistic processing of the responses and the kind of feedback generated by activities with linguistic processing of the responses by means of NLP tools. The session included an overview of general aspects in NLP, so that teachers could grasp what is actually happening with the learner's response, and the information that serves as a basis for the system to make judgements on the performance of the learner.

11.2.1.4 Managing AutoTutor

This presentation introduced AutoTutor as a tool for authoring and managing ICALL activities. The presentation included a detailed explanation of the response specification process, the ReSS, as well as the use of the graphical interface of the authoring tool. Additionally, teachers were trained on the steps required to upload an activity to Moodle and the way it would be accessed and used by learners. A description of available teacher tracking facilities was also included.

11.2.2 Material creation process

The material creation process consisted of two kinds of activities on the teacher side. One of them was individual work. Teachers were expected to devote a weekly average of two to four hours to design and create the materials. The second kind of activity was a monthly group meeting, which the three participating teachers and I attended. Teachers were encouraged to communicate with each other and with me through the material creation process.

11.2.3 Application of materials in class

The use of materials in class consisted of the following steps:

Presentation of the experiment to learners We helped teachers set up an introductory class including a presentation of the experiment, a leaflet with instructions on how to use AutoTutor as a learner and we had the learners respond to a profile questionnaire. This included the creation of a Moodle account in the server hosting the AutoTutor module.

Use of materials with learners Teachers guided learners through the classes and required them to access the Moodle server were AutoTutor activities were available together with other activities planned for the experiment's work.

Conclusion of the experiment with learners We attended the class considered by teacher as the one closing the experiment, and exchanged opinions with learners.

11.3 Results

This section describes the results and the product generated by the participants in the experiment. We include a description of the materials created by the three teachers and a description of the use made by learners of the corresponding activities.

11.3.1 Authored materials

We analyse the authored materials from four different perspectives. First we will comment on the integration of the materials in the course programme. Second we will describe the input data used by the teachers in their materials. Third, we will analyse the activities in terms of the TAF. And, fourth, we will analyse them in terms of the RIF.

11.3.1.1 Integration in course programme

Materials created by Teacher 1

T1 designed a series of activities based on the work plan reflected in Figure 11.11, the whole work plan can be found in Annex H. The topics handled are: Physical and Chemical Changes, Sublimation, Reactants and Products, Catalysts, Stoichiometry, and Yields and Rates. As the figure shows, the materials include three types of activities: *folder* activities, a kind of portfolio, laboratory activities, and CALL activities. Different types of activities are combined and integrated within and across sessions. The total number of sessions foreseen in this workplan is eleven. Some sessions expand over several hours, which means that they are hold through more than one 50-minute class, or that part of the work is done at home.

As shown in Figure 11.10, the CALL materials created by T1 include a total of ten activities and are unevenly distributed among four of the topics in the workplan. Out of the ten activities, seven of them are plain CALL activities with immediate feedback not requiring NLP processing – multiple choice, matching, and cross-word activities authored with Hot Potatoes. The other three are AutoTutor activities – with an icon representing a human head with its mouth open.

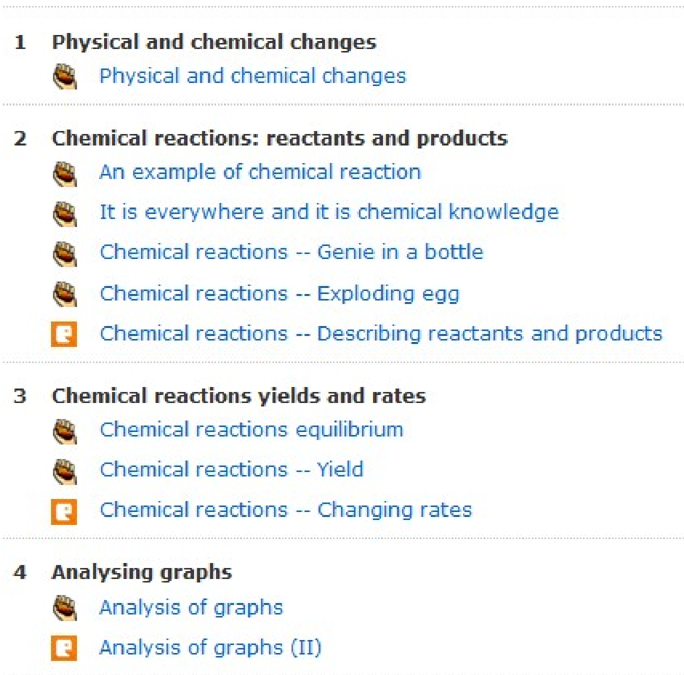


Figure 11.10: Fragment of the workplan designed by T1 to include ICALL and CALL materials in the course sessions on chemical reactions.

CHEMISTRY Physical and chemical changes. Chemical reactions			
Activities and writing (folder) / Laboratory	Computers activity	SCIENCE 3rd of ESO	
Sessions	**Meeting point**	**Time**	**Resources**
Session 1: Class and Computers activity SCIENCE TOPIC: Physical and chemical changes • Teacher explains at the blackboard the differences between physical and chemical changes. Students write down to include in their folders. • Then the whole class watch the Video: "Physical and chemical changes" and put in common some examples to tell them apart. • Then the class meet in computers room in order to do the Hot Potatoes activities related to the video. - Physical and chemical changes Match	In class In computers room	2 hours	Computers Internet connection
Session 2. Laboratory SCIENCE TOPIC: Physical changes. Sublimation. • The sublimation of Iodine.	Laboratory	1 hour	Laboratory instrumental Iodine

Session 3, 4. In class and in computers room.
SCIENCE TOPIC: Chemical reactions. Reactants and products.
• Teacher explains at the blackboard examples of chemical reactions, types and the differences between reactants and products. Students write down to include in their folders.
• Next students do the Hot Potatoes activities:
  - Chemical reactions Quiz
  - Chemical reactions Cloze
  - Chemical reactions Cross File
	In class
	In computers room
		2 hours	Computers
			Internet connection

Figure 11.11: Fragment of the workplan designed by T1 to include ICALL and CALL materials in the course sessions on chemical reactions.

Materials created by Teachers 2 and 3

T2 and T3 designed a unit of work consisting of three CALL activities to be used as supplementary materials by learners to review the Present Simple. The actual activities as seen in the Moodle course page are shown in Figure 11.12. T2 and T3 did not integrate the activities created in their regular course programme; they were used as an addendum to review the topics handled during the course. The activities are planned to be used in one, at most two sessions giving learners time to work at their own pace.

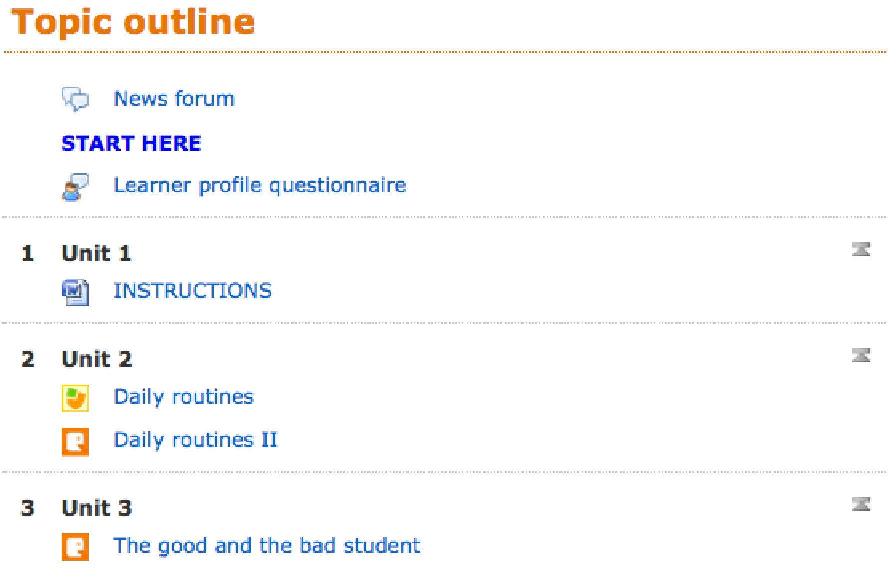


Figure 11.12: Overview of the materials created by T2 and T3 for the ESL courses.

The figure reflects three different pieces of materials: an MS Word document with instructions on how to proceed, a JClic activity including several fill-the-gap and matching exercises to warm up[6], and two AutoTutor activities.

11.3.1.2 Input data

All the CALL/ICALL materials created during the experiment resulted in HTML pages including several sources of input to the learner. T1 used videos, images, equations, metalinguistic information, and metalinguistic rules in the input data. T2 and T3 used images and videos, and metalinguistic information, but no metalinguistic rules – more details below. Annex H includes screen captures of the AutoTutor activities that reflect the kind of input data included. The actual activities can be accessed at http://iceee.barcelonamedia.org/courses.

[6]This activity was not created by them but taken from one of the available repositories on the Internet.

11.3.1.3 TAF characterisation

This section is an overall presentation of the TAF characteristics of the activities created by the teachers. For a detailed characterisation of all the AutoTutor activities using the TAF analysis see Appendix H.

Activities by T1

The three AutoTutor activities created by T1 are: *Chemical reactions – Describing reactants and products*, from now on A1-T1, *Chemical reactions – Changing rates*, A2-T1, and *Analysing of graphs (II)*, A3-T1. Though the activities created by T1 are CLIL activities, we comment them under the perspective of the frameworks presented in Part III of the thesis, the Task Analysis Framework. The complete analyses of the activities are in Annex H.

The activities created by T1 are writing activities that focus on formal aspects of the subject area, Chemistry, and that take the foreign language as a means to communicate the knowledge in the area. They can be considered activities focusing on meaning and form, and they can be classified as pre-communicative language learning. When we presented the TAF and the RIF, we based them on approaches and proposals derived from research and practice in communicative approaches to language teaching. Though CLIL is not mentioned there, it is accepted commonly in the FLTL and the SLA communities that CLIL is part of the communicative approach to language learning – in fact, it is seen as "natural development of communicative approaches" (Pérez-Vidal, 2009: p. 6).

Each of the ICALL activities of T1 has a pedagogical goal related with the subject Chemistry which is parallel to the linguistic goal. Thus, A1-T1 expects learners to practice both the reading and interpretation of chemical equations and the use of passive and active voices. A2-T1 expects learners to practice also the reading and interpretation of chemical equations, but including the yields, that is the amounts of chemical product obtained given the mole ratio; this activity expects learners to use conditional structures. Finally, A3-T1 expects learner to be able to read graphs reflecting chemical or physical processes, and to be able to use comparison structures.

It is arguable whether the cognitive processes of these activities are directly related to real-life processes, but they are pedagogically relevant to consolidate the concepts worked on in the lab and classroom activities. In terms of the responses, the three of them require limited production responses. A1-T1 has seven items to respond to, A2-T1 has eight, and A3-T1 includes also eight items. All items in all activities present graphical hints (either chemical equations or graphs images) and the input data always include examples and a metalinguistic explanation of the responding procedure.

Activities by T2 and T3

T2 and T3 created two AutoTutor activities: "Daily routines II", from now on A1-T2/3, and "The good and the bad student", from now on A2-T2/3. Both of them are writing activities that can be classified as communicative language practice. Both activities focus on practising the present simple tense and the use of frequency adverbs. A1-T2/3 has as an additional goal: to foster the use of time expressions such as *at a quarter to eight*.

The cognitive processes in both activities have similarities with real-life processes, such as describing one's habits, for instance, when visiting a doctor, or describing third-party actions. In terms of responses, both of them require limited production responses. A1-T2/3 includes eight questions: The first three require learners to write one of their morning routines, that is, they are required to write three in total; from the fourth to the sixth learners are required to write three of their afternoon routines, one at a time; and the last two questions expect them to write two evening routines.

11.3.1.4 RIF characterisation

This section is an overall presentation of the RIF characteristics of the activities created by the teachers. For a detailed characterisation of all the activities using the RIF analysis see Appendix H.

Activities by T1

Relationship between input and response The three activities created by Teacher 1 follow a quite similar pattern in terms of relationship between input and response. The three of them include a number of supporting input data ranging from metalinguistic formulas to response samples. One of them, A1T1 includes a video as a warm-up activity, to activate certain concepts and vocabulary in learners' brains.

The activities created by T1 respond to Task Type I (see 7.2.2.1). They present a narrow scope in the relationship between input and response. Though they do not follow the pattern ask-about-X and use-expression-Y in the instructions, as that task did, they include a simple image and a short text that try to elicit the expected knowledge and language. The information to be processed is short, and the kind of response required as a reaction is also short.

As for the directness of the relationship, all responses in all items of these three activities present a direct relationship to the input data in terms of topical and language knowledge. Both of them are restricted by the images and the short texts included in each item.

Thematic and linguistic contents First we characterise the activities in terms of thematic contents, that is, in terms of the entities and the relations expected in the response. As for entities, the items in the three activities created by T1 include references to chemical elements and compounds (*water, lithium, carbon dioxide...*), chemical processes or reaction drivers (*electrolysis, oxidation, catalyst...*), units of measure (*moles, kg...*), and a few other chemistry or other common concepts such as *concentration* (of a substance) or *time* (as variable that determines the evolution of chemical reactions). In terms of relations, the two activities on chemical reactions only refer to the generation of products. The third one includes a more varied range of relations: evolution of concentration, state changes, temperature changes, and evolution of the degree of solubility.

As for the linguistic content of the responses, we start with the functional contents: The first two activities are basically oriented to elicit pieces of language to describe chemical reactions alternatively including either the chemical process or its yield. These relations determine the lexical contents expected to express the process that leads to the transformation of the reactants into the products (*produce, give, yield, generate, form...*). A difference between these two activities is that while the former expects the learner to describe the reaction by mentioning the chemical materials in it, the latter expects them to produce the sentence as if they were explaining what is or has been the yield of a reaction *if you* have X moles of one of the reactants or products. The third activity has again a different type of functional content and is oriented to elicit graph interpretations from the learner. In this case, lexical contents are determined highly by the graphs and partly by the input data in form of text for each question.

From a syntactic point of view, each of the activities has a very different purpose. A1-T1 is oriented to practising the active and passive voice; A2-T1 is oriented to practising conditional sentences; and A3-T1 is oriented to practising comparative structures. Learners are expected to use properly the word order in the relative simple sentences that they are expected to use. These activities have no specific pragmatic goals, except that of observing social norms in class work, and in terms of graphology learners are expected to produce well-formed full sentences.

11.3.1.4.1 Activities by T2 and T3

Relationship between input and response. The two activities created by Teachers 2 and 3 follow a similar pattern in terms of relationship between input and response. Both of them include mainly visual input data and only some metalinguistic guidance as part of the instructions. The information to be processed is short, and the kind of response required as a reaction is also short. Both of them are similar to the Type I activity (see 7.2.2.1). They present a narrow scope of the relationship between input and response. However, they differ in two respects: on the one hand they have no linguistic input data at the level of item/questions, on the other hand they ask the learner to resource to their own experiences or habits to complete the responses.

As for the directness of the relationship, the responses to the items of these two activities present a relatively direct relationship to the input data in terms of topical and language knowledge. Both of them are restricted by the images and the short texts included as input data in each item. However, there is an intention to elicit subjective topical knowledge, which opens the door to unexpected language forms. These two activities provide some room for creativity, since the instructions allow learners to respond on the basis of their personal experience.

Thematic and linguistic contents. We start with the characterisation of the activities in terms of the entities and the relations expected in the response. As for entities, the items in the two activities created by T2 and T3 include references to a number of semantic fields. Responses to A1-T2/3 include references to objects and people learners come across with in their daily routine: from an *alarm clock* to a *guitar* through a *bus*, the *school*, his *father* or *mother*, and so on. T2 and T3 intend to restrict this range by using some images as input data. In terms of relations, A1-T2/3 responses include references to commuting *go to*, *return*, or *go back*, meal related actions such as *have breakfast/lunch/dinner*, or leisure activities or hobbies such *playing* + MUSICAL INSTRUMENT, *swimming*, and so on.

As for A2-T2/3, the expected responses contain references to entities that are objects and people related to school and study: *exams*, *homework*, *teacher*, *classmate*, and so on. In terms of relations, the expected responses include references to actions such as *listen to*, *pay attention*, *talk*, *chew (gum)*, and so on.

As for the linguistic content of the responses, we start with the functional contents. Both activities are oriented to elicit pieces of language to describe routine actions and habits in two contexts: personal habits for A1-T2/3, and socially well and badly marked habits among learners in school for A2-T2/3. The wording of the input data for each question opens the lexical choices for learners to respond. For instance, Item 1 in A1-T2/3 reads "Write one of your morning routines". This opens the response to literally any habit or activity the learner does from the time he or she wakes up until midday. However, through visual input data and the work done in class teachers expect learners to produce words such as *school*, *comic*, *book*, *homework*, *watch TV* and so on. Similarly, for A2-T2/3, Item 1 reads "What does a perfect student do in class?" Again, this opens a space for the learner to respond using his or her own experience. Once more, through images and a tag cloud used as input data teachers expect learners to produce words such as *have a shower*, *surf the Internet*, *newspaper*, *read*, *chat*, *friends*, *computer games*, and so on.

From a syntactic point of view, both activities have a similar purpose: They are both oriented to practice the production of simple sentences using the present tense, frequency adverbs and time expressions. However, for the responses to A1-T2/3 learners are expected to be the subject of the sentences (*I, me, myself*), while for the responses to A2-T2/3 learners are expected to refer to third parties, good and bad students (*he*, *she*, *the good/bad student...*). Learners are expected to use correctly the word order and include all elements in simple sentences in the present. As it happened with T1's activities, they have no specific pragmatic goals, except that of observing social norms in class work, and in terms of graphology learners are expected to produce correct full sentences.

11.3.2 ReSS-based specifications by teachers

This section describes and analyses the ReSS specifications generated by teachers by using the authoring tool and the proposed response specification procedure.

11.3.2.1 Specifications by T1

We describe the response specifications generated by T1 for the activity *Chemical Reactions – Describing reactants and products*. The qualitative analysis is based on Item 1 of this activity; the quantitative analysis focuses on all items.

Qualitative analysis

Figure 11.13 shows the Response Components, the variants and the strings, and the RC sequences of the response specifications for Item 1 of activity A1-T1. Figure 11.13a presents rectangles in darker grey marking RC, rectangles in lighter grey marking Variants, and rectangles in white representing Strings. Figure 11.13b presents RC sequences.

As shown in Figure 11.13a, T1 specifies five RC for this activity's response:

- RC A contains the reactants of the chemical equation, *salt* and *water*;
- RC B contains the different ways in which the transformation process can be expressed. B1 contains the forms for the responses in active voice, and B2 the ones for those in passive voice;
- RC C contains the products of the equation, *chlorine*, *hydrogen* and *sodium hydroxide*;
- RC D contains the different ways of expressing the chemical process causing the reaction, *electrolysis*;
- and RC E is a period, for responses to end with a period.

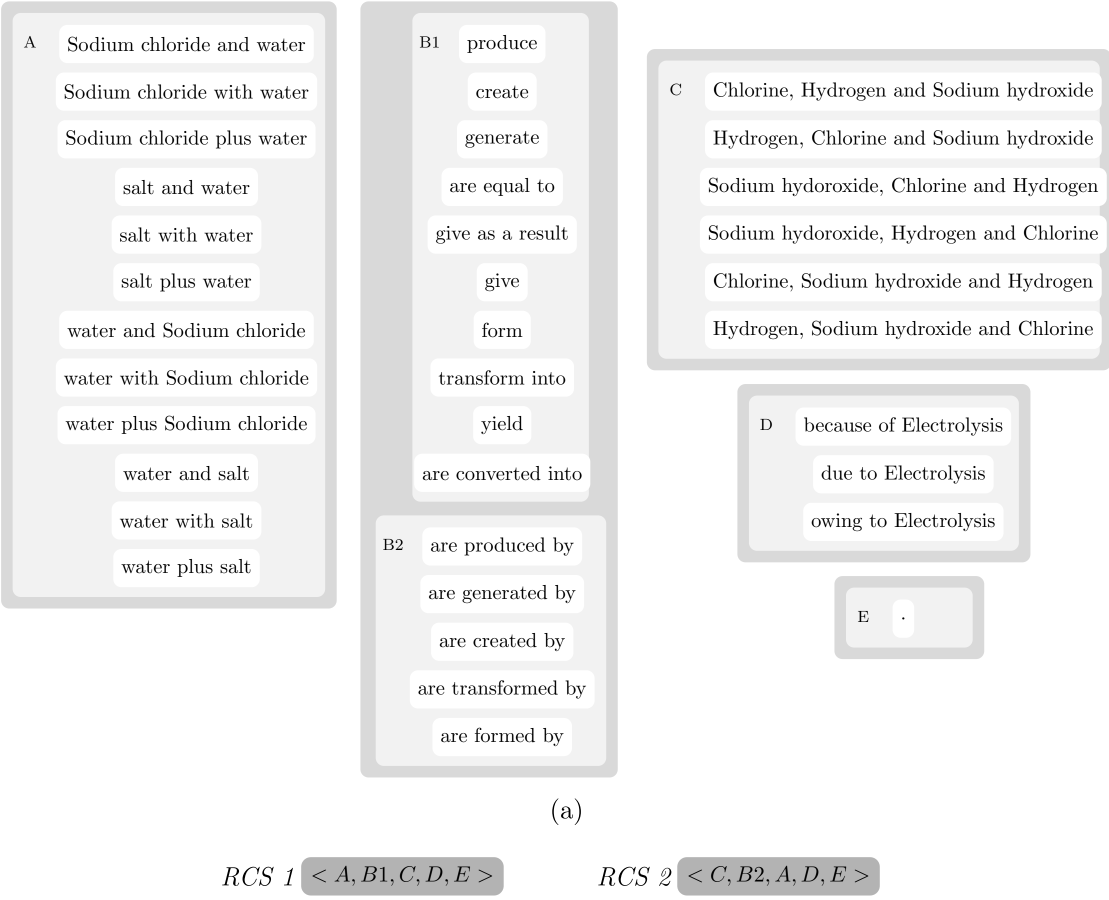
The partitioning of the response into RCs corresponds with the notions of entities and relations of the RIF, while the partitioning into variants corresponds to different linguistic materialisations of the corresponding concepts. For instance, the relation *X produces Y* can be expressed in active, B1, or in passive form, B2.

As for the RC sequences in Figure 11.13b, there are four: RCS 1 and RCS 2 correspond to responses in active voice, while RCS 3 to RCS 4 correspond to responses in passive voice. Note the differences between RCS1 and RCS2, and RCS 3 and RCS 4 are only RC E, the response component containing the period. This is the only way teachers could use to indicate that they accepted responses both with and without a period at the end of the sentence.

In Appendix H, we can see that the three activities generated by T1 follow very similar patterns in terms of ReSS characterisation.

Quantitative analysis

Table 11.2 shows the complexity of the response specifications for all the items in A1-T1 in terms of Response Components, Variants, Strings, RC sequences and total number of sentences generated. For a relatively low number of RC, Variants, Strings, and RCS, the total number of well-formed correct responses that can be generated can add up to 3240. We also see that for some of the items the number of generated sentences can be very low. The total number of sentences generated for Item 7 is 15 times lower than that generated for Item 1.



(b)

Figure 11.13: ReSS specifications for Item 1 in A1-T1.

11.3.2.2 Specifications by T2 and T3

This section describes the response specifications generated by T2 and T3 for the activity "Daily Routines II". Again, the qualitative analysis is based on Item 1; quantitative analysis focuses on all items of this activity. Figure 11.14 shows the Response Components, the variants and the strings, and the RC sequences of the response specifications for Item 1 of the activity A1-T1. The correspondence between colours and ReSS elements is the same as before: darker grey rectangles mark RC, lighter grey ones mark Variants, and white rectangles represent Strings. Figure 11.14b presents RC sequences.

As shown in Figure 11.14a, T1 specifies five RC for this activity's response:

- RC A contains the subject of the routine action (I, the speaker);
- RC B contains the different activities that the speaker might perform: *have a shower*, *read a book*, *have breakfast/lunch/dinner*, and so on. The distinction between B1 and B2 is, according to the teachers, that they expect learners to produce sentences including frequency adverbs with the actions in B1 but not in B2, as can be seen by comparing RCS 4 and RCS 8.

Activity	RC	Var	Str	RCS	Sent
A1-T1 Item 1	5	6	38	2	3240
A1-T1 Item 2	5	6	29	2	2808
A1-T1 Item 3	5	6	28	2	1296
A1-T1 Item 4	5	6	30	2	1728
A1-T1 Item 5	5	6	31	2	2592
A1-T1 Item 6	5	6	29	2	624
A1-T1 Item 7	5	6	23	2	216
Avg.	5	6	30 (±5)	2	1786 (±1145)

Table 11.2: Number of RC, RC Variants and Strings, RCS and total number of generated sentences for activity items in A1-T1.

- RC C contains the period, a punctuation sign;
- RC D contains frequency adverbs;
- RC E contains four variants, each of them containing different strings necessary to construct time expressions in combination with the RCs F and G;
- RC F contains cardinals that can be used to build time expressions containing the hour name: one, two, three...;
- RC G contains the preposition *at*, to build time expressions such as *at twenty to eleven*;

The response specifications for this Item foresee five RCs: the one referring to the speaker, the one referring to the action (the routine), the one referring to the frequency (adverbs), the ones referring to particular times of the day, and the punctuation sign. However, the RC referring to the time of the day is split into three blocks, a decision that reduces enormously the need to write down possible times of the day.

As for the RC sequences, there are 17 different possible orderings of the blocks. We can group them into four main patterns: One that generates sentences including B1 actions with no frequency adverbs, such as $< A, B1, G, E1, F, C >$; one that includes sentences with B2 actions and no frequency adverbs, such as $< A, B2, G, E3, F, C >$; and the other two are the same groups but allowing frequency adverbs, such as $<A, D, B1, G, E3, F, C>$, $<A, D, B2, G, E3, F, C>$.

In Appendix H, we can see that the two activities generated by T2/3 follow similar patterns in terms of ReSS characterisation.

Quantitative analysis

Table 11.3 shows the complexity of the response specifications for all the items in A1-T2/3 in terms of Response Components, Variants, Strings, RC sequences and total number of sentences generated. The total number of well-formed correct responses that can be generated can add up to 21376, which is much higher than the 3240 possible correct responses that could be generated with the specifications of the item analysed for A1-T1. As before, for some of the items the number of generated sentences can be very low in comparison. The total number of sentences generated for Items 1–3 is roughly 3.5 times the number of sentences generated for Items 7–8.

Activity	RC	Var	Str	RCS	Sentences
A1-T1 Item 1–3	7	11	60	17	21376
A1-T1 Item 4–6	7	11	50	17	11246
A1-T1 Item 7–8	7	11	48	17	6632
Avg.	7	11	53 (±5)	17	14026 (±6664)

Table 11.3: Number of RC, RC Variants and Strings, RCS and total number of generated sentences for activity items in A1-T2/3.

(a)

RCS 1 $< A, B1, G, E1, F, C >$

RCS 2 $< A, B1, G, E3, F, C >$

RCS 3 $< A, B1, G, F, E2, C >$

RCS 4 $< A, B2, C >$

RCS 5 $< A, B2, G, E1, F, C >$

RCS 6 $< A, B2, G, E3, F, C >$

RCS 7 $< A, B2, G, F, E2, C >$

RCS 8 $< A, D, B1, C >$

RCS 9 $< A, D, B1, G, E1, F, C >$

RCS 10 $< A, D, B1, G, E3, F, C >$

RCS 11 $< A, D, B1, G, E4, F, C >$

RCS 12 $< A, D, B1, G, F, C >$

RCS 13 $< A, D, B1, G, F, E2, C >$

RCS 14 $< A, D, B2, G, E1, F, C >$

RCS 15 $< A, D, B2, G, E3, F, C >$

RCS 16 $< A, D, B2, G, E4, F, C >$

RCS 17 $< A, D, B2, G, F, E2, C >$

(b)

Figure 11.14: ReSS specifications for Item 1 in A1-T1.

11.3.2.3 Overview of the complexity of response specifications

Table 11.4 shows the expansion from ReSS specifications to a set of well-formed correct responses to a particular activity. While the average number of RC, Variants, Strings and RC sequences remains low, the average number of sentences generated for each activity (that is, for the whole items in it) ranges from a few hundreds to 14000.

Activity	RC	Var	Str	RCS	Sentences	
A1-T1	5 (±0)	6 (±0)	30 (±5)	2 (±0)	1786	(±1145)
A2-T1	6 (±0)	11 (±1)	28 (±3)	11 (±3)	373	(±231)
A3-T1	8 (±1)	11 (±2)	31 (±6)	4 (±1)	496	(±310)
A1-T2/3	7 (±0)	11 (±0)	53 (±6)	17 (±0)	14026	(±6664)
A2-T2/3	4 (±0)	7 (±0)	47 (±4)	4 (±0)	554	(±49)

Table 11.4: Total number of Response Components, Variants and Strings per question in the ICALL activities authored by teachers.

11.3.3 Use of materials by learners

The materials were used in class by teachers as described in Section 11.2.3. This produced a series of actions on the learner side, which we describe in this section.

11.3.3.1 Use of materials by T1

Teacher 1 used the materials that he created with the three groups in the third year of obligatory secondary education described in Section 11.1.2.2. All groups started using the created materials on the week of the 2nd to the 6th of May 2011, and had the closing session of the experiment on the week of the 30th of May 2011 to the 3rd of June 2011. The initial and closing sessions were held in my presence. The rest of classes and work was carried out either in the classroom or at home.

Learners in the three groups working with T1 did not manage to complete all the activities proposed. As a result, (almost) no learner responses were collected for Activity A3-T1. Table 11.5 reflects the activity generated by learners during learning experiences with the AutoTutor activities in terms of attempts. All attempts are counted at the same level: that is, first attempts are not counted separately. Of course, some learners might have attempted only once to respond and others up to six or seven times, but we are not focusing now on this kind of learner behaviour.

As shown in the table, all groups start by submitting a higher number of responses to be corrected and the number of attempts decreases as the learners progress in the individual questions in that activity. This is particularly reflected in the decreasing number of attempts as we progress from Q1 to Q8 (or Q7) for the three groups.

11.3.3.2 Use of materials by T2/3

Teachers 2 and 3 used their materials with the four different groups of the first and second year of obligatory secondary education described in Section 11.1.2.2. All

Activity	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
3A								
A1-T1	62	76	54	66	40	45	51	–
A2-T1	68	49	5	6	11	5	6	13
A3-T1	0	0	0	0	0	3	0	0
3B								
A1-T1	140	88	79	87	52	74	59	–
A2-T1	55	34	17	21	11	9	12	12
A3-T1	9	3	0	0	0	0	0	0
3C								
A1-T1	201	77	69	76	34	56	45	–
A2-T1	3	4	2	1	1	1	2	1
A3-T1	0	0	0	0	0	0	0	0

Table 11.5: Response attempts by learners in 3A, 3B and 3C working with materials generated by T1.

groups worked with the materials between the 20th and the 23rd of June 2011. I was not present in the initial and closing sessions.

Most learners in the four groups working with T2 and T3 did manage to complete all the activities proposed. Table 11.6 reflects the activity generated by learners during their learning experiences with the AutoTutor activities in terms of attempts. Again, for all groups the number of attempts decreases as the learners progress in the individual questions in that activity.

However, in this table an additional fact is observed: That questions 2 and 3 (Q2 and Q3) had no attempts for the four groups. This is a consequence of misunderstanding or unclear instructions of the activity to be performed. Learners were expected to write three of their morning routines in three different text areas: one in Q1, another one in Q2, and a third one in Q3. In the end, they all provided the three of them in the text area for Q1; so, after that, teachers required them to go on with Q4 and then write only one routine per text area.

11.3.4 Quality and usefulness of the feedback

In this section we analyse the response attempts and the feedback provided by AutoTutor for one of the activities performed by three of the seven groups of learners presented. We will analyse in further detail the data generated by groups 1A and 2B to activity A1-T2/3, and by group 3B to activity A1-T1. There is no particular reason in choosing these three groups or activities except that: (i) each of them was led by a different teacher, T2 lead 1A, T3 led 2B and T1 led 3B; and (ii) the three selected activities contained a higher (or the highest) number of response attempts compared with the activities performed by other groups in the same levels.

AutoTutor's activity player presented feedback to learners in two steps (since it is based on the feedback architecture presented in Section 8.3). They were expected to submit each response for correction twice. In the first submission they checked

Activity	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
1A								
A1-T2/3	90	0	0	37	36	16	12	17
A2-T2/3	28	44	23	15	–	–	–	–
1B								
A1-T2/3	103	0	0	28	30	17	11	13
A2-T2/3	25	16	23	16	–	–	–	–
2A								
A1-T2/3	43	0	0	11	9	8	8	13
A2-T2/3	35	31	28	19	–	–	–	–
2B								
A1-T2/3	143	0	0	47	43	23	19	24
A2-T2/3	29	32	23	20	–	–	–	–

Table 11.6: Response attempts by learners in 1A, 1B, 2A and 2B working with materials generated by T2/3.

grammar and spelling. In the second their response was checked in terms of task-specific thematic and linguistic contents.

11.3.4.1 Criteria for the evaluation of feedback

The goodness of the feedback for both correction steps was evaluated following similar criteria. Tables 11.7 and 11.8 show how these were respectively applied to the evaluation of the spell and grammar checking functionality, and to the task-specific content and language checking functionality.

As shown in Table 11.7, the Response/Feedback pair no. 1 is validated as "False". It shows a sentence in which the word *Iron* was marked as being incorrectly written in capital letters. Though this might be a correct feedback under circumstances it is not in this context, because T1 asked his pupils to use capital letters to write element and compound names. Though the system had a device to detect unknown words or minor writing differences between the teacher's specifications and the standard criteria, this did not work here because the dictionary information required the word *iron* to be written in lower case.

Response/feedback pair no. 2 is validated as "True". It shows a response for which a valid feedback message was generated, meaning a message that really detects an error, and the messages is consistent with the activity's correction criteria. Response/feedback pair no. 2 is a simple spelling error message were **breaksfast* was written instead of *breakfast*.

Response/feedback pair no. 3 is validated as "Bad". It shows a response for which a misleading feedback message was generated, meaning a message that detects a real error, but the explanation that it provides has little or nothing to do with the explanation of the error according to the activity's correction criteria. As shown in the table, the message warns the learner of the use of two noun phrases in a sentence, while the error is a spelling error in the second word, *kisten*. Though this second

No.	Response/Feedback pair	Validation
1	R: Iron with Oxygen produce **Iron** (III) oxide due to oxidation. F: Use lower case for this word.	FALSE
2	R: I have **breaksfast** at eight o'clock. F: Check if this is a spelling error.	TRUE
3	R: **I kisten**[1,2] to music at eight o'clock. F: < 1 > A sentence cannot start with two noun phrases. Check whether there is a mistake. <2> Does the word *kisten* contain a spelling error?	BAD

Table 11.7: Validation strategy for the evaluation of feedback quality in terms of spelling and grammar.

error is actually detected in a second message, the question is the learner can be misled by reading all the information.

In Table 11.8, we describe the application of these same criteria for the evaluation of task-specific thematic and linguistic contents. The Response/Feedback pair no. 1 is marked as "False". The message warns the learner of having written a verb, *salt*, in an unexpected form. However, the word *salt* in this context should clearly not be interpreted as a verb. Other errors in the response, such as the writing of **hidroxide* instead of *hydroxide*, and the ambiguous reading of *salt*, as a verb and as a noun, cause this misbehaviour.

No.	Response/Feedback pair	Validation
1	R: Hydrochloric acid and Sodium hidroxide produce **salt** and water because of Neutralisation. F: Check if this is a verb that should have a different form.	FALSE
2	R: Carbon **oxide** and water are produced by Glucose and Oxygen because of Respiration. F: This part of the answer does not correspond to any part of the answers stored by the system. Please check!	TRUE
3	R: Carbon dioxide and water **are** formed **by** Glucose **plus** Oxygen plus Carbon dioxide because of Photosynthesis. F: These words do not correspond to the expected response. Please check if they are needed.	BAD

Table 11.8: Validation strategy followed for the evaluation of system performance in terms of task-specific content and language.

Response/feedback pair no. 2 is validated as "True". It shows a response for which a valid feedback message was generated, meaning a message that really detects an error, and the messages are consistent with the activity's goals. Response/feedback pair no. 5 shows a content error. The expected reactant is *Carbon dioxide*, not $^{\#}$*Carbon oxide*.

Response/feedback pair no. 3 is validated as "Bad". It shows a response for which a misleading feedback message was generated. The response includes the compound *Carbon dioxide* as a reactant and describes the process as "Photosynthesis". Both are incorrect. The expected response does not include *Carbon dioxide* in the reactants, and the process should be "Respiration". These differences cause the system to mark the words *are*, *by* and *plus* as not needed, when in fact they are required. Though the complete feedback message correctly identifies the problems with the second occurrence of *Carbon dioxide* and with *Photosynthesis*, the feedback is misleading.

11.3.4.2 Feedback to step one in the correction process

Table 11.9 shows the number of feedback messages generated for each activity and group that conveyed false information (False), true information (True), or bad information (Bad). A fourth column in Table 11.9 shows the response submissions attempts that did not obtain a response due to an error during the communication between the client and the server. Finally, the last two columns show the total sum (Sum) of feedback messages generated and the total number of response submissions (*Att*empts). Note that we count attempts and feedback messages separately, since one response attempt can obtain more than one feedback message.

In terms of grammar and spell checking the system's performance was excellent: for activity A1-A1T2/3 it obtained a 97% and 99% of accuracy, while for activity A1-T1 it obtained a 96%. Looking at the figures globally, 97% of the messages issued by the system respond to real errors, 2% of the messages generated in the responses informed about errors that did not exist, and 1% of the messages informed about real errors whose explanation was not adequate. Only two of more than 1000 response submissions resulted in a server communication error, which explains the imperceptible percentage of submissions where there was a server connection error.

	False	True	Bad	Conn. F.	Sum	Att.
1A-A1T2/3	0	206	6	1	213	205
2B-A1T2/3	1	293	0	1	295	294
3B-A1T1	29	682	2	0	713	537
Total	30	1181	8	2	1221	1036
Percentage	2	97	1	0	100	–

Table 11.9: Goodness of ICALL feedback in spell and grammar checking.

11.3.4.3 Feedback to step two in the correction process

Table 11.10 shows the same data we just described for the first correction step but for the second correction step. The performance of the system here is reasonably good. It presents an accuracy of 76% and 63% for A1-T2/3 for groups 1A and 2B, and of 60% for A1-T1 for group 3B. The messages generated by the responses submitted by group 3B result in false errors 3% of the time, and in incorrectly diagnosed errors 13% of the time. For the other two groups the number of false and incorrectly diagnosed errors is always zero.

However, a striking figure is the number of responses that were not submitted for this second correction step, independently of the group: 24% for 1A, 37% for 2B, and 27% for 3B. These figures reflect a difficulty in learners to grasp that the correction process consisted of two steps, first spell and grammar checking and then exercise specific language and content checking.

	False	True	Bad	Not sub.	Sum	Att.
1ESO-A1T2/3	0 (0)	176 (76)	0 (0)	55 (24)	233	205
2ESO-A1T2/3	0 (0)	200 (63)	0 (0)	118 (37)	318	294
3ESO-A1T1	3 (0)	591 (60)	133 (13)	266 (27)	993	537

Table 11.10: Goodness of ICALL feedback in activity specific language and content checking.

11.3.4.4 Error analysis of the system's performance

Error analysis of feedback messages in spell and grammar checking

The only false error found for activity A1-T2/31 in group 2B was due to the system not being able to detect the use of the construction "$^{??}$*At morning* I wake up at seven o'clock". Though apparently this expression can be used in British English, it is obsolete and very rare in modern English.

The false errors found for activity A1-T1 in group 3B are related to differences between the teacher's criteria in the use of capital letters in chemical nomenclature compared to the default system's behaviour. This led to the system marking the use of capital letters for the words *Iron* and *Hydrochloric (acid)* as incorrect, while the teacher required specifically to use them. There was also an error due to a lack of coverage in the lexicon. Though the feedback generation system includes a strategy for allowing words unknown to the system if the teachers includes them in specifications as correct, it does not have a strategy to handle words from the domain, that is, Chemistry, that are not included in the lexicon.

As for errors that received an incorrect explanation, most of them are related to structures that seem to have a concordance problem between subject and predicate or within a noun phrase, but in fact hide other kinds of errors. This is the case for sentences like **Ibrush my teeth* or **I ususally brush my teeth at five past eight*, where the system complains about the sentences starting with two nouns. Though the message in itself is literally true, the problem is clearly that the unknown word detection heuristics of the system have overgenerated in considering "Ibrush" and "ususally" a noun. One of the bad messages is related with word formation rules, particularly one that favours the writing of "*carbondioxide" instead of the correct one, "carbon dioxide".

Error analysis of feedback messages in activity specific checking

The feedback messages containing incorrect information about real errors are quantified and typified in Table 11.11.

Type of information in message	Abs. frequency
Unexpected word(s).	32
Unpexted word(s) and missing word(s).	30
Should the verb be in a different form?	16
A conjunction marker is missing.	14
Check if word(s) is (are) needed.	11
Check if preposition/particle is wrong.	10
This word seems to be in the wrong form.	6
Keywords are relevant but something is wrong.	5
Check if 'and' suits better/is needed.	4
This list requires more items.	2
The word(s) are in the wrong position in the sentence.	2
Start word with capital letter.	1
Total	133

Table 11.11: Frequency and nature of the messages that yielded wrong explanations in responses that included real errors or deviations.

A detailed analysis of the technical reasons that cause the system to behave like this will not be pursued in this thesis. However, in order to give the reader an impression of the kind of system behaviour that we are alluding too, we will briefly comment on one particular message generated for a real learner response to Question 1 in Activity A1-T1. For this question and activity, the expected response was one that had to include a message as the one reflected in (70), but the learner response in (71) generated the feedback in Figure 11.15.

(70) Sodium chloride and water form Sodium hydroxide, Hydrogen and Chlorine because of Electrolysis.

(71) #Sodium Chlorine plus water produce Sodium *Hidroxide* with *water* and Chlorine *duet* to Electrolysis.

The response contains four words that imply an error. Two of them are spelling errors, namely *Hidroxide* instead of Hydroxide and *duet* instead of *due*. The other two are content errors. The learner wrote *Sodium Chlorine* as a reactant instead of Sodium chloride, and she or he also wrote *water* as a product, when *Hydrogen* was expected.

The feedback generated by the system is shown in Figure 11.15. While all errors are actually detected and marked (see messages 3, 4, 5 and 6), there is a number of messages, namely 1, 2, 7, 8 and 9, that contain information that is misleading or incorrect. Some of the individual messages are false errors, others are bad messages. In 3 the learner is told that *water* is not expected, while in 8 he or she is told that the expression *water and* should be moved to the beginning of the sentence. The

1. **Sodium** Chlorine plus water produce Sodium Hidroxide with water and Chlorine duet to Electrolysis
   This part of the answer contains relevant keywords but something is wrong. Please check!

2. Sodium Chlorine plus water produce **Sodium** Hidroxide with water and Chlorine duet to Electrolysis
   This part of the answer contains relevant keywords but something is wrong. Please check!

3. Sodium Chlorine plus **water** produce Sodium Hidroxide with water and Chlorine duet to Electrolysis
   This part of the answer does not correspond to any part of the answers stored by the system. Please check!

4. Sodium Chlorine plus water produce Sodium **Hidroxide** with water and Chlorine duet to Electrolysis
   This part of the answer does not correspond to any part of the answers stored by the system. Please check!

5. Sodium Chlorine plus water produce Sodium Hidroxide with water and **Chlorine duet** to Electrolysis
   This part of the answer does not correspond to any part of the answers stored by the system. Please check!

6. Sodium **Chlorine** plus water produce Sodium Hidroxide with water and Chlorine duet to Electrolysis
   Word order! This expression is in the wrong position. Locate it further in the back of the sentence.

7. Sodium Chlorine plus water **produce** Sodium Hidroxide with water and Chlorine duet to Electrolysis
   Word order! This expression is in the wrong position. Locate it further in the front of the sentence.

8. Sodium Chlorine plus water produce Sodium Hidroxide with **water and** Chlorine duet to Electrolysis
   Word order! This expression is in the wrong position. Locate it further in the front of the sentence.

9. Sodium Chlorine plus water produce Sodium Hidroxide with water and Chlorine duet **to** Electrolysis
   Check if this preposition or particle is wrong.

Figure 11.15: Feedback messages generated for a learner response that include some misleading information.

question is the system expects *water* to be in the first part of the sentence, since this is an active, not a passive sentence, but then, in order, to mark the error it chooses to consider an error the occurrence of water that is further away from its position, and that is not combined with *Sodium chlorine*.

11.3.4.5 Learner uptake

As explained in the background chapters, validating the technical feasibility of an ICALL system is not enough to be able to assert that the feedback that the system is providing is useful. In this section we analyse the degree to which learners did profit from the feedback obtained. To do so, we analyse the percentage of resubmissions that include a change in the response that can be correlated with the feedback message, but only for those feedback messages that were considered valid in the previous section.

11.3.4.5.1 Criteria for the evaluation of learner uptake

Tables 11.12 and 11.13 reflect the criteria under which learner take up is evaluated. We will consider as no uptake taking place at all, those resubmissions that do not reflect a modification of the response in respect to one or more specific feedback messages, as well as those feedback presentations that were not followed by a resubmission. The first kind of ignored feedback is reflected in the triads response/feedback/resubmission no. 1 in each of the two tables. In Table 11.12 triad no. 1 shows a warning about punctuation marks being ignored. In Table 11.13 no. 1 shows a content error (oxide instead of hydroxide) being ignored.

No.	Response/Feedback/Resubmission triad	Reaction
Step one: spell and grammar checking		
1	R: I sometimes eat a snack at six **o'clock**	Ignored
	F: Sentences should end with a punctuation mark.	
	Resub: I sometimes eat a snack at six *o'clock*	
2	R: **Carbon oxide with water are** produced by Glucose and Oxygen because of Respiration.	Profit
	F: Subject and verb do not agree.	
	Resub: *Carbon oxide and water are* produced by glucose and Oxygen because of Respiration.	
3	R: Sodium chloride and water produce Sodium **hidroxide** and Hydrogen and Chlorine.	No profit
	F: Check if this is a spelling error.	
	Resub: Sodium chloride and water produce Sodium *hydoxide* and Hydrogen and Chlorine.	
4	R: Carbon dioxide plus **Water** produce Glucose plus Oxygen because of Photosynthesis.	Alternative
	F: Check if this word should be in lower case.	
	Resub: Carbon dioxide plus Hydrogen oxide produce Glucose plus Oxygen because of Photosynthesis.	

Table 11.12: Analysing learner take up on the basis of changes in resubmissions for the correction of spelling and grammar errors.

Triads no. 2 in each table show examples of resubmissions showing there was take up either in terms of spell and grammar checking, as in Table 11.12, or in terms of activity specific language checking, as in 11.13. Triads no. 3 show examples of resubmissions showing there was take up, but this did not lead to a correct response. This is again exemplified in in terms spell and grammar checking in Table 11.12, and in terms of activity specific language checking in 11.13. Finally, triads no. 4 exemplify resubmissions showing how learners opt for rephrasing their response. For instance, an upper case warning in Table 11.12 ends up in the learner replacing Water with Hydrogen, while a content-related warning in Table 11.13 ends up in the learner replacing *have a sandwich* with *listen to music*.

No.	Response/Feedback/Resubmission triad	Reaction
	Step two: activity specific checking	
1	R: Salt and water forms Sodium **Oxide**[2], Hydrogen and Chlorine because of Electrolysis.	Ignored
	F: This part of the answer does not correspond to any part of the answers stored by the system. Please check!	
	Resub: Salt and water produces Sodium *Oxide*, Hydrogen and Chlorine because of Electrolysis.	
2	R: I usually go to **the** school at a quarter to nine.	Profit
	F: Your response is correct, but check if this is relevant.	
	Resub: I usually go *to school* at a quarter to nine.	
3	R: Salt and water **produces**[1] sodium **oxide**[2] and **hydrogen**[3] and **chlorine**[4] because of Electrolysis.	No profit
	F: $<1>$ Check if the form without the s-ending is better. $<2>$ This part of the answer does not correspond to any part of the answers stored by the system. Please check! $<3,4>$ Capitalise this word.	
	Resub: Salt and water *produces* sodium *oxide* and *hydrogen* and *chlorine* because of Electrolysis.	
4	R: I have **a sandwich** at a quarter past five.	Alternative
	F: This part of the answer does not correspond to any part of the answers stored by the system. In addition, there is something missing.	
	Resub: I *listen to music* at a quarter past five.	

Table 11.13: Analysing learner take up on the basis of changes in resubmissions for the correction of activity specific language and content.

11.3.4.5.2 Quantitative analysis of learner uptake

Figures 11.16 and 11.17 show the percentage of resubmissions that given a previous feedback show the learner ignores the feedback (grey column, labelled 'Ignored'), profits from the feedback (green column labelled 'Profit'), does not profit from the feedback (yellow column, labelled 'No profit'), or uses an alternative structure or expression to respond.

Figure 11.16 shows these percentages for feedback messages generated for spelling and grammar. The percentage of messages ignored is the highest value both for groups 1A and 3B, 57.4% and 47.6% respectively. For group 2B, the percentage of ignored messages is 9.1%.

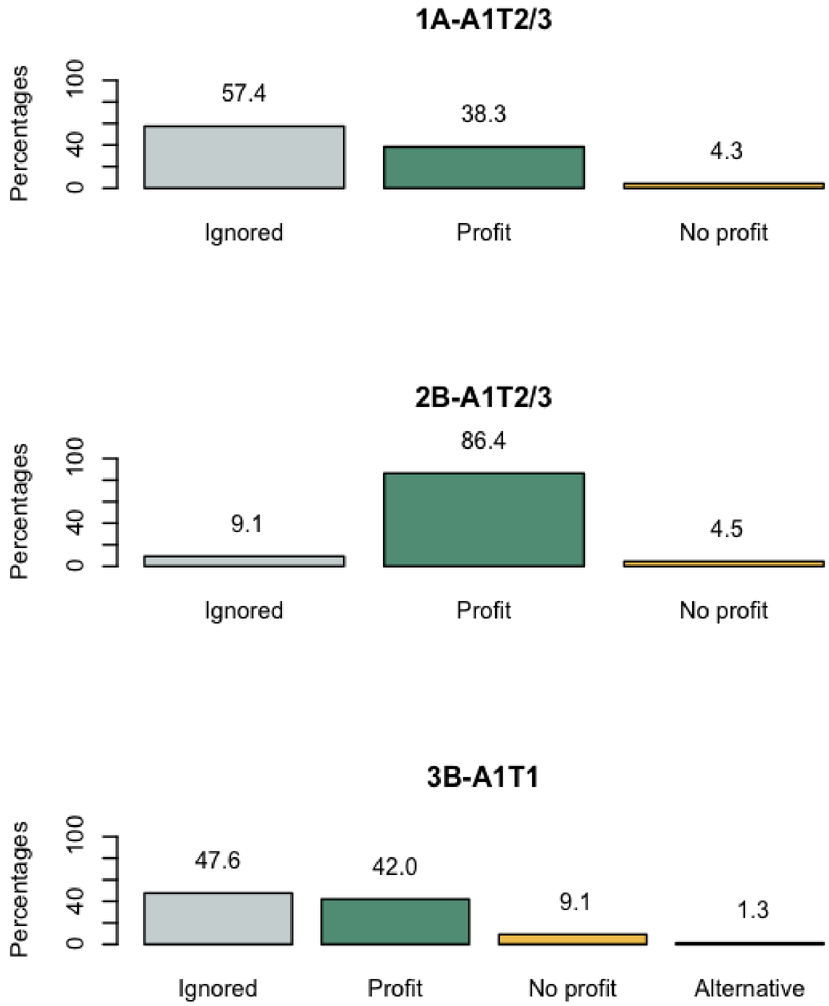


Figure 11.16: Learner uptake for valid feedback messages to spelling and grammar errors.

As for the percentage of resubmissions showing profit from the feedback, the highest value is the one obtained for group 3B, namely 86.4%. Groups 1A and 3B obtain respectively percentages of 38.3% and 42%. As for the percentage of learner resubmissions showing that the learner does not profit from the messages, these are 4.3%, 4.5% and 9.1% respectively for groups 1A, 2B, and 3B. Group 3B includes a small percentage of resubmissions, 1.3%, that reflect that learners decided to go for a response using a structure different than the original one.

Figure 11.17 shows the resubmission percentages showing whether learners profited from the feedback messages generated for activity-specific language and content. Again, groups 1A and 3B show a higher percentage of resubmissions where the learner ignored the messages, respectively 60.7% and 64.5%. In contrast, for group 2B this figure goes down to 41.4%.

As for the percentage of resubmissions showing profit in the subsequent response, the highest value is for group 2B, 41.4%. Groups 1A and 3B present respectively values of 26.2% and 23.4%. As for the percentage of learner resubmissions showing that the learner does not profit from the messages, these are 13.1%, 25% and 11% respectively for groups 1A, 2B, and 3B. Group 3B includes a small percentage of resubmissions, 1.1%, that reflect that learners decided to go for a response using a structure different than the original one.

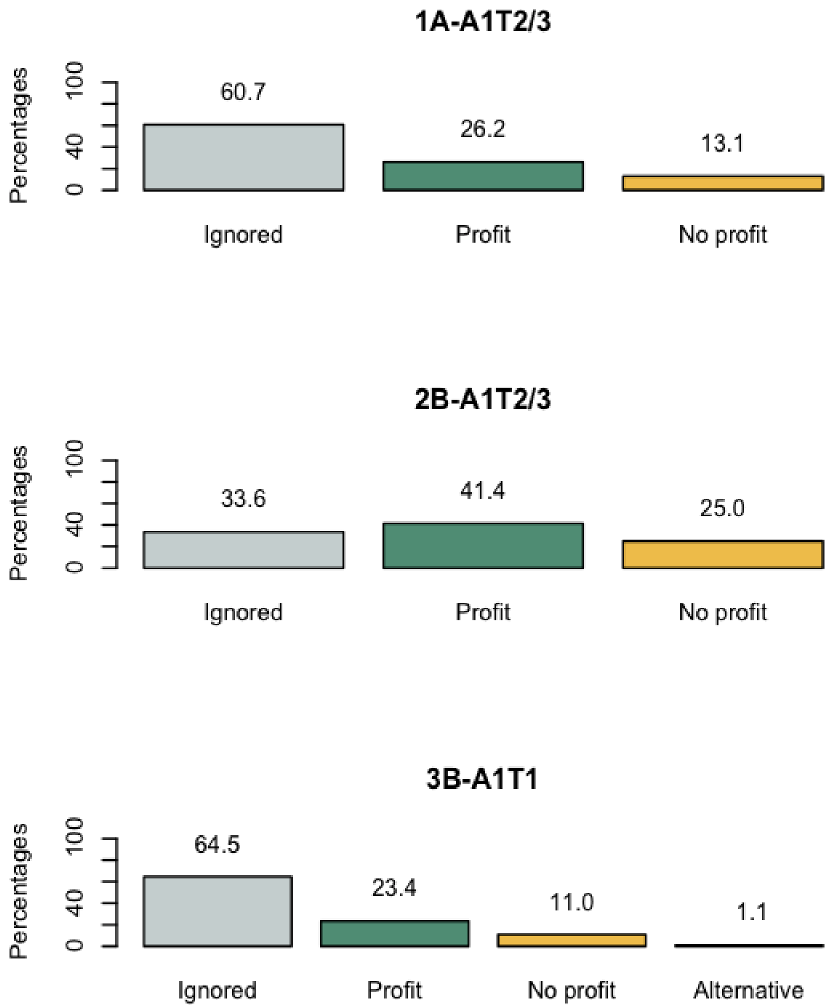


Figure 11.17: Learner uptake for valid feedback messages to activity specific language and content errors.

11.4 Discussion

11.4.1 Teacher perspective

After the experiments we conducted a series of interviews and passed a questionnaire to the participant teachers. The following paragraphs summarise teachers' views and opinions.

About system's feedback

The three participant teachers believe that the feedback provided to learners by AutoTutor helped their learners significantly improve (5 in a scale from 1 to 5) their spelling. They also thought that it helped learners improve the sentence structure of their language quite a bit (4).

While T1 believed that AutoTutor's feedback helped learners improve their grammar a lot (5), T2 and T3 felt it only helped them moderately (3). In terms of vocabulary, T1 thought that AutoTutor helped their learners quite a lot (4), and he also thought that it helped them improve their topical knowledge, that is, Chemistry related knowledge. T2 and T3 did not value whether AutoTutor's feedback helped learners at all in terms of vocabulary or contents.

About the effects on their teaching process

The responses provided by T1 and T2/T3 in this respect have a very different nature. According to T1 the materials that he created with AutoTutor allowed him to create activities combining language skills with Chemistry contents in a more integrated manner. T2 and T3 said that the experience of participating in this project helped them reflect on "the error-making process of our students", as well as to become more conscious of the steps their learners have to go through to provide correct sentences.

About the added value of experimenting with AutoTutor

The three participant teachers agreed that using AutoTutor facilitated the correction task, helped them by highlighting the mistakes most frequently made by their learners, and helped in general improve the correction and feedback process. Moreover, the three of them thought that it was a positive influence on their learner's motivation. In particular, T1 emphasised that through AutoTutor activities he engaged many learners in the interest for looking for a solution and devote time to it, rather than the frequent behaviour of getting the right response as quick as possible and forgetting about the task.

In addition, T1 believed that this experiment helped him integrate language and ICT activities in class, as well as integrate language learning with the learning of other subjects. As for T2 and T3, they thought that using AutoTutor allowed greater student autonomy.

About the material creation process

The three participant teachers thought the activity authoring process was "very slow", particularly certain aspects of the interface made it too clumsy and unfriendly. The three of them made specific improvement requirements during and after the experimentation. T2 and T3 thought that the activity creation process was a continuous trial-and-error process oftentimes requiring a redesign of the concept.

During the material creation process, four issues were identified as creating most of the problems. The first was related to the difficulties in producing response specifications. Sometimes teachers tried to produce as part of the specifications a series of incorrect responses that they expected learners to produce for such an activity. Sometimes they divided responses in too many or too few components or variants, and that made the automatic correction of the responses more difficult.

A second problem was related to the usability and the proper operation of the graphical interface and its functionalities. Bugs and errors were found and repaired during the experiment.

A third problem was the actual understanding and definition, by the teachers, of the task to be performed. Teachers were not always conscious that we were asking them to produce materials for themselves and for their learners, not materials accomplishing some sort of pedagogical characteristic that we could be particularly interested in. Probably, they are not used to think about the production of materials as such a thorough process, including the specification and prediction of responses.

Finally, teachers often had problems determining whether or not a particular activity was really suitable to be corrected using NLP strategies. They were often told that activities allowing for open responses were not suited, and they understood what it meant. However, teachers are probably not used to restrict the range of possible responses to the required extent, since their presence in class or the assumption of a human correcting the responses allow for a greater flexibility. For instance, T2 and T3 modified up to four times the general concept and definition of their unit of work.

11.4.1.1 AutoTutor's feedback compared to teacher feedback

As a means to compare the system's performance to what teachers would have said to learners for a particular response, we provided T2/3 with a list of responses that were submitted by T2's learners. We asked them to correct those responses as if they had to give them back to learners with formative feedback, including all but not more information than they would usually include.

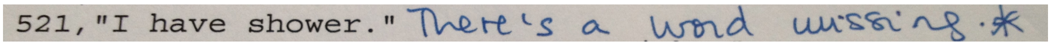
With the teacher's correction, we went through each of the feedback messages provided by the system for each response and annotated whether that feedback message was included or not in the teacher's manual correction. Since the feedback generated for one response could include more than one message, we analysed each of the messages for a response individually.

Criteria for the comparison of system feedback with teacher corrections

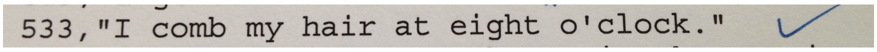
Our comparison process considered three different options: the teacher agrees, the teacher does not agree, or the teacher cannot agree because there was no submission.

The last option is included for responses not submitted to the second correction step.

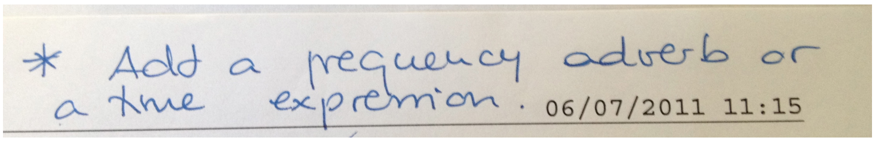
Figure 11.18 exemplifies the teacher's manual corrections for two different responses, namely Figures 11.18a and 11.18c. In Figure 11.18a, the response "I have shower" gets two messages from the teacher: One that reads "there's a word missing", and a second one that reads "Add a frequency adverb or a time expression.", referenced to using an asterisk (whose explanation is shown in Figure 11.18b). In Figure 11.18c, the response "I comb my hair at eight o'clock." gets a tick from the teacher indicating it is correct.



(a)



(b)



(c)

Figure 11.18: Learner responses corrected manually by T2.

With these teacher corrections, our system register files can be enhanced with a column that states whether the teacher's correction explicitly or implicitly agrees with the system's feedback, as shown in Table 11.14. Note the validation is carried out taking into account the teacher's correction and the system generated feedback as a whole. Thus the message in the first row, Attempt 7218 Step 1, agrees with the teacher comments because the error marked by the teacher is marked in Step 2, see the second row.

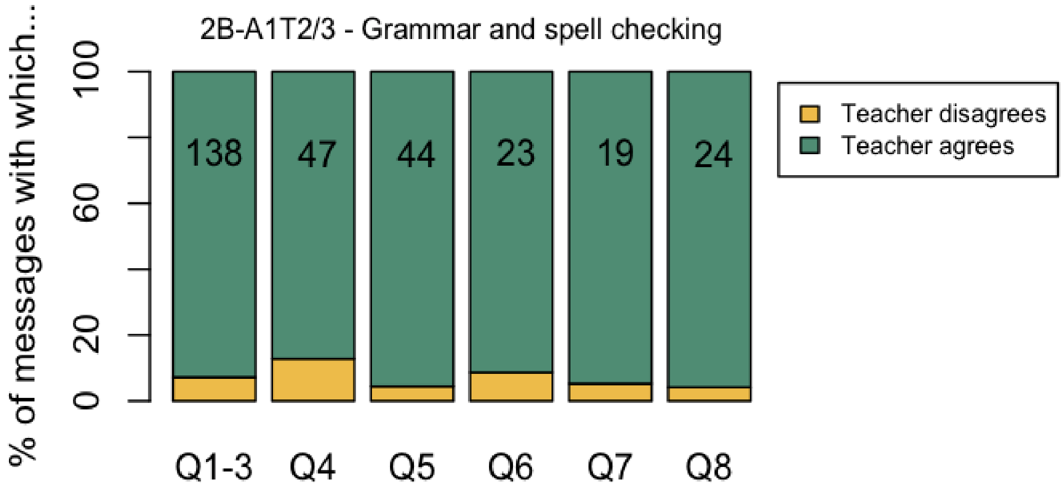
Att.	Step	Feedback	Agr.
7280	1	No spelling or grammar errors.	1
7280	2	Check if a determiner 'a' or 'an' is missing.	1
(...)			
7319	1	No spelling or grammar errors.	1
7319	2	Correct answer	1

Table 11.14: Excerpt of the system-teacher comparison register.

As for the second annotation, third and fourth row regarding Attempt 7319, the agreement annotations are implicit. The teacher does not explicitly produce the messages given by the system, but we assume a tick is compatible with these two messages. Simply machines and humans provide messages in a different way, of course.

Quantitative analysis of system-teacher agreement

Figure 11.19 shows the distribution of positive and negative agreement between the system and the teacher's feedback in percentages for the two different correction steps. The figures are presented separately for each question in activity A1-T2/3, and the figures in the central part of the bars are the absolute number of feedback messages compared for each particular question.



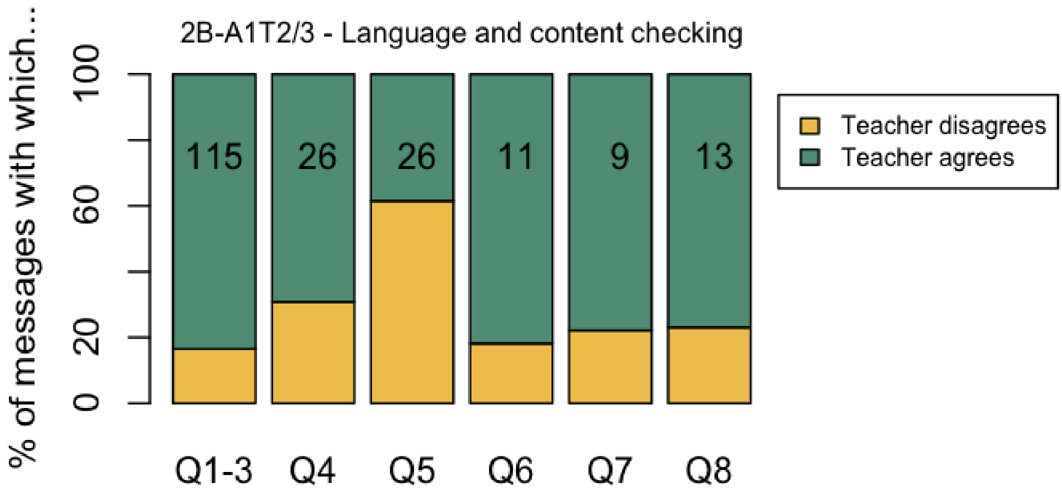


Figure 11.19: Agreement between system's feedback and teacher comments.

For the first correction step, the overwhelming presence of green reflects a large degree of agreement between the system's messages and the teacher's comments.

This would indicate that in terms of general spell and grammar checking teachers and AutoTutor produce very similar kinds of remarks. It must be said, however, that disagreement does not necessarily mean that the system incorrectly indicated an error. It can be the case that the system indicates that a sentence is not ended with a punctuation mark, but the teacher does not mark that as an error, because, as the teacher said in a post interview, she is then focusing on other pedagogical goals. Or it can be the case that the system does not identify an error that the teacher does identify.

As for the agreement between teacher and system in the second correction step, green is still the colour of predominance, but the proportions are less favourable to the system than they were before. This suggests again a reasonable degree of similarity between system and teacher remarks.

However, in Question 5 the percentage of disagreements is clearly much larger than the percentage of agreements. When looking into the details we observe for this activity that learners, a group of them, decided to go for more creative responses, and this had clearly the effect that they were not properly handled by the system, while a teacher could never mark them as incorrect, given the activity's specifications.

With respect to this second correction step, we observed that there was a number of responses that could not be validated because learners did not submit the responses for the second correction step. We exclude the data because it would amount to evaluating the friendliness of the system rather than agreement; we do not intend to diminish its importance, particularly since the number of responses not submitted to task-specific correction ranges between 25% and 50%.

11.4.2 Learner perspective

Though the experimentation process included a final session with learners in which they were required to respond a satisfaction questionnaire, we only obtained them for groups 3A, 3B and 3C, that is, Teacher 1 groups. We will briefly summarise their opinions in the following paragraphs.

About the feedback provided by AutoTutor

Figure 11.20 the responses of the learners to the question "How helpful to you were the [AutoTutor] exercises providing language feedback?", on a scale from 5 (very helpful) to 1 (not at all). According to their responses, 53% of the learners thought it was quite useful (4), 38% of them though it was moderately useful (3), and 8.8% thought it was not very useful (2).

Moreover, learners were asked specifically about what aspects of their knowledge were better supported by the feedback provided by AutoTutor. About their spelling, 2.9% learners found that AutoTutor helped them a lot (5), 35.3% quite a lot (4), 47.1% moderately (3), 8.8% not very much (2), and 5.9% not at all (1). About their grammar, 2.9% of learners found that AutoTutor helped them a lot (5), 47.1% quite a lot (4), 38.2% moderately (3), and 11.8% not very much (2).

Learners were also asked whether AutoTutor's feedback helped them in understanding of meaning, in general: 5.9% of learners found that AutoTutor helped them

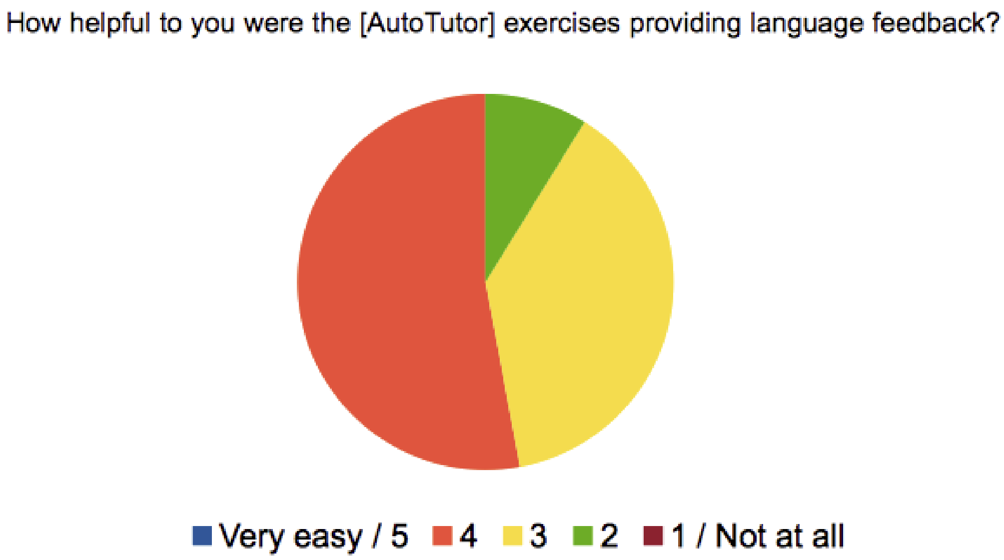


Figure 11.20: Satisfaction of learners with AutoTutor activity feedback.

a lot (5), 38.2% quite a lot (4), 44.1% moderately (3), and 11.8% not very much (2). Finally, we asked learners whether it helped them in the organisation of their writing: 8.8% of learners found that AutoTutor helped them a lot (5), 47.1% quite a lot (4), 41.2% moderately (3), and 2.9% not very much (2).

11.4.3 Research perspective

From the research perspective, there are three main discussion topics we would like to pay attention to, all of which are related to the second goal of the thesis, namely to assess the feasibility of developing a technology and a methodology for the authoring of meaningful and useful ICALL activities by teachers. First, there is material creation, approached as a process and as a product. Second, there is the use of the materials in class, particularly how teachers handle the limitations of the generated feedback and how learners profit from the learning experience. Last, there is the technical shortcomings, particularly in terms of NLP.

11.4.3.1 Material creation: the process and the product

During the material creation process Teachers 2 and 3 had more difficulties than Teacher 1 complying with the experiment instructions. Our three most important requirements were that they authored activities that could be responded to using one sentence, included input data to which the learners could resource to, and fit in the course programme – ideally used as supplementary work to class activity.

Integration of materials in class work

As for the integration in the course programme, Teacher 1 did conceive a working plan for the third trimester that integrated AutoTutor, and other CALL activities.

The work plan, as shown in Section 11.3.1 and in Annex H, included detailed plans of the work to be completed in class, in the laboratory, and in computer rooms. In contrast, Teacher 2 and Teacher 3 created an independent unit of work that was consistent with the topic and grammar syllabus of the course, but that was only used as a reinforcement after the regular classes had ended. An important factor in this difference is that Teacher 1 taught this CLIL course for the first time, and it was also the first time that it was offered in the school too. So he had an additional motivation to create the materials.

However, the EFL teachers had a reason that might explain this too. In a post interview with them, Teachers 2 and 3 stated that they worked hard to develop an activity that allowed their learners to work on specific writing micro-skills in activities where responses could include a minimum space for creativity. For instance, they thought about activities where learners could be asked to describe people or objects using more than one adjective, as an exercise that would prepare them to write longer descriptive texts.

With this goal in mind Teachers 2 and 3 considered up to five different pedagogical concepts, but for several reasons these did not work. One of the reasons was that the responses they expected were too open and they could not find a way to make them more restricted and still useful. They were ready to reduce the spectrum of possible answers, but did not want to end up producing a comprehension or a picture description activity. This might explain why they spent more time producing pedagogically interesting activities that could be corrected automatically with a tool like AutoTutor, than in integrating them in their class work. In our opinion, this reflects their desire to fully comprehend the possibilities and limitations of this new technology before using it generally in class.

Input data to support learners in responding

As described in Section 11.3.1.4, the types of activities respectively created by T1, on the one side, and T2 and T3, on the other side, follow similar characteristics. While T1 activities are less open to linguistic creativity, T2 and T3's are. The input data that T1 provides in his activities often determine the contents of the response, while the input data that T2 and T3 provide tend to suggest possible contents to the response. Particularly in A1-T2/3 response variation is much larger than in the other activities, as we saw in Section 11.3.2.3.

Nonetheless, as we saw in Section 11.3.4, this difference in response variation does not affect system behaviour in terms of feedback generation. Neither in terms of the correction of spelling and grammar, which is to a certain extent comprehensible, nor in terms of activity specific language and content checking, which really counters our expectations. The figures in Table 11.10 suggest that the behaviour is even better for A1-T2/3, though this would certainly require further investigation.

Maybe the only evidence we find that supports our expectations, based on the TAF and RIF analysis, that A1-T2/3 was more open than any of the activities created by T1 is the bad performance of the assessment for Question 5 in Figure 11.19.

Response type in authored activities

The three participant teachers created activities that could be responded to with one sentence. This was mostly ensured by the strict specification options in the interface, as well as by our explanations during the introductory sessions. T2 and T3 had a harder time producing an activity with such a short response, as they stated, since they really wanted to author an activity to prepare learners for the production of a longer text. As a result, they created A1-T2/3, which asks learners to state three of their morning routines, three of their afternoon routines, and two of their evening routines. Each of the individual routines has its own text area in terms of HTML form. This had the advantage that, as they did, they could use the same response specifications for three different questions, which simplifies the authoring process.

However, it has at least two disadvantages. One of them did show up in the phase during which materials were used in class. All learners responded to their first three morning routines in the same text area. After the appropriate explanations they did it following the instructions for the other five questions, so a better wording of the instructions might help. The second disadvantage did not show up, namely that a learner could "successfully" complete the activity by providing the same response in the first three questions, a second response in the second three questions, and a third response in the last two. This latter disadvantage could be overcome technically, but it would requires a more detailed analysis in terms of usefulness, since often one action performed in an experiment like this is later on not reproduced, or not worth reproducing, in real life.

The process of designing and authoring the activities

In light of the results and the monitoring performed, we think that the design process is comparable to the design process of other kinds of learning activities or learning activities using other supports or technologies. Though there was a considerable learning curve, our three participant teachers said the amount of time devoted and the complexity in terms of additional work was reasonable. The added value of ICALL materials is that, if strategically planned and designed, they can be re-used and iteratively improved with relative ease.

As for the aspects that make the process difficult, we advert that it is, not only because of the not very friendly proof-of-concept interface used to introduce all the activity-related information, but also because of the effort of conceiving good, meaningful activities to be automatically corrected. This latter ability requires experience with the limits and the capabilities of NLP-enhanced technologies.

The ReSS as a natural way of specifying responses

As the experiment progressed, teachers were able to manage with agility the concepts Response Component, Variant and Response Component Sequence. That responses require certain "concepts" to be correct, that these concepts can be expressed in different words, and that not all words combine equally is a notion teachers possess innately as speakers of a language. It was natural in them to rely on these abilities to apply the ReSS.

However, it was also realised very quickly that the level of concreteness required by AutoTutor, which was not able to expand response specifications on the basis of linguistic similarities, was too painstaking and time-consuming. As a result, T2 and T3 authored activities for which the same response specifications could be used to correct different responses (see Figures H.29 through H.35 in Appendix H). This led to results that were not always satisfactory, though we draw from it two positive conclusions: First, the fact that teachers understood enough about the specification process to find out strategies to reduce their manual work. Second, we see it as demand for the inclusion of functionalities in the authoring tool to make the teacher's work easier.

In this respect, one should consider how more abstract levels of representation of criteria for correctness could be successfully used. Such a goal would imply the experimentation with novel NLP approaches capable of enhancing the comparability of linguistic expressions regarding their meaning, such as Rich Textual Entailment. As well, it would probably require the elaboration of frameworks for the specification of criteria for correctness more abstract than the ones we proposed.

The understanding of ICALL/NLP

An aspect that we identified during the material creation process is that teachers eventually develop a sense of what NLP and ICALL actually are: their possibilities and functionalities. For instance, in a discussion regarding why the system would not accept as correct the sentence "I sometimes read manga at three o'clock", while, in accordance with teacher specifications, it accepted the sentence "I sometimes read a comic at three o'clock", they rapidly saw that the system they were using lacked any kind of semantic analysis or meaning inference functionalities.

Similarly, in developing the specifications in terms of Response Components and Response Component Sequences, they soon came up with the need to develop sort of lists of lexical items that could be used for more than one activity. For instance, T2 and T3 wanted to create lists of adverbs expressing frequency in positive sentence (always, sometimes, usually...) or the list of adverbs expressing frequency in negative sentences (rarely, never, seldom...). As for T1, he wanted to create a list of the verbs that can be used to express that reaction is taking place (forms, produces...). Interestingly, these are strategies that computational linguists use in hand-written rule-based systems.

In this respect, the work done by T2 and T3 allowed them to take the possibilities of AutoTutor to the limit with relatively satisfactory results. They managed to develop activities to practice pre-writing activities, as they intended. However, they were not totally satisfied with them since the pedagogical concept left too little room for creativity, precisely because they had to be corrected automatically. This is a call for the use of NLP techniques to allow for an expansion of the responses with little supervision. These would go in the direction of applying something as translation memories but for the purpose of correction, or in the direction of using meaning inference techniques such as dictionary-based synonym detection or rich textual entailment – in line with the work by Bailey and Meurers (2009).

11.4.3.2 Materials used in class

Activity management by teachers

Teachers seemed to be comfortable with the management of activities during their use in class or with the learners at home. While they did not use the student tracking facilities, they did interact a lot with their learners in order to learn about the quality of the feedback they were obtaining.

As for T1, this interaction supposed several changes in the response specifications that went from correcting minor typos to the enhancing the response spectrum with new words originally not considered by him. Since AutoTutor allowed him to modify the response specifications without the need to upload the materials into the course management system again, he used this interaction to improve the automatic correction functionality of his materials from one class to the next (recall he used the materials for three different groups). He even used this possibility in real time, while using the materials in class. However, one should consider here the effects of having the same group of learners experiencing different system behaviours.

T2 and T3 did not use this possibility to improve their response specifications during the use of the materials, though they were using the materials for only two sessions with each group and the time elapsed between each session was very short – they used the materials with all the groups in roughly three days.

Activity use by learners

The empirical data we analysed to measure the profit that learners did take from using AutoTutor activities was the evidence reflected in response resubmissions that could be explained by learners taking into consideration a feedback message. The one salient thing in this analysis is that there was also a large percentage of messages that did not show any evidence, or, as we put it, they were ignored: from 9% to 65%, depending on the correction step. Nonetheless, we also saw there is a reasonable percentage of messages that resulted in uptake, varying from 23% to 86%, again depending on the group and correction step. Finally, there was a lower percentage of feedback messages that produced a wrong reaction on the learner side ranging from 4% to 25%.

As for ignored messages, they do not necessarily imply there was no take up at all. That is, when looking into the learner activity registers, one sees learners that after a couple of submissions do not re-submit one last time the response if the error is a punctuation, spelling, or grammar error. They probably consider these minor errors, or simply mistakes, and not errors, following Corder (1974)'s terminology.

As for the messages that misled the learners, inspecting the logs suggests that often the messages are expressed using a language that is opaque and too technical for learners this age. The satisfaction questionnaire revealed that, on a scale from 1 to 5, 38% learners considered that interpreting the system's feedback was neither very useful nor little useful (3), and 8.8% considered it was little useful (2).

There is a parallel observation on the teacher side, since the three of them expressed this concern at several points during the experiment. In fact, T3 insisted repeatedly during the experiment to have access to the files with the canned messages so that she could rewrite and improve them. For several reasons this was not possible, but this confirms that customising an ICALL system will be something more than introducing response specifications. It will require aspects such as determining the type of feedback to be generated, summative and/or formative, details to be included in the feedback (localisation, explanation, revealing solutions or not), or types of language used in the feedback (more or less metalinguistic, no text at all for certain types of error, etc.). These are indeed research lines that will be of interest not only to ICALL, but to CALL in general, and that further support Levy (1997: p. 42)'s argument in favour of CALL systems offering teachers much more control over the learning materials and, more generally, system behaviour.

11.4.3.3 The limits of AutoTutor's NLP-based feedback

Variation as a problem for NLP

One of the important goals of the experiment was to achieve the production of ICALL activities that were computationally feasible, meaning activities whose answers where within the range of answers that could be generated by the system using teacher specifications. This range of responses was partially ensured by the way the experiment was designed, requiring teachers to work with activities whose responses were limited production responses with a narrow and direct relationship between input and response. However, during the analysis of the results we did observe responses that, most of the time being correct, did not match with any of the possible responses generated by the system according to teacher specifications.

As we saw in Figure 11.19, for the second correction step there were between 18% and 25% of the messages that did not show agreement between the system and the teacher, except for one of the questions for which this went up to 61%. This supposed a total of 50 out of 318 messages for the whole activity (that is, adding up the messages generated for all the questions in it). Inspecting the logs, we observe that 48 out of this 50 messages correspond to sentences showing language and content variation with respect to the initial teacher expectations.

Among the responses, we find instances of lexical variation as the ones in (72). The responses that would have been accepted are the ones below in italics. In both cases it is a question of synonymy, at least in the context.

(72) a. I return home at five o'clock.
I COME BACK *home at five o'clock.*
b. I start LESSONS at nine o'clock.
I start school at nine o'clock.

The response in (73) exemplifies syntactic variation. Teachers were not expecting coordinated sentences. This could be because they wanted learners to express only one action per question, or simply because they did not come up with the idea. In any case, the system was not able to handle properly this variation, because it simply did not make use of any techniques that could use the given information, the given response specifications, to reach it.

(73) I brush my hair and my teeth at a quarter to nine.
I brush my hair at a quarter to nine., or
I brush my teeth at a quarter to nine.

Another type of variation is the one reflected in (74), at the level of functional knowledge. As we can see "before having dinner" is probably as good a time reference as a concrete time is, such as, "a quarter to eight". However, this was neither expected according to the response specifications, nor the teacher real expectations.

(74) I watch TV before having dinner.
I watch TV at TIME EXPRESSION.

A last type of variation observed is the one reflected in (75), at the level of topical knowledge. These are three different attempts of the same learner trying to have the response accepted. Independently of the fact that the responses might contain errors or aspects to be commented on, the concept "practising a sport" did not appear among the expected responses.

(75) a. I always train waterpolo.
b. I always train my sport.
c. I go to the sports center.

Note the different types of variation obey to either variation in terms of linguistic knowledge or variation in terms of topical knowledge. Linguistic variation raises a technological question, that is, the challenge to develop NLP strategies that with a minimum amount of specifications handle the maximum range of responses. As for the topical variation, not only does it raise it a technical question, but also a pedagogical one. Even if the system were able to handle this kind of variation, aspects such as the pedagogical goal of the activity should be considered.

Automatic analysis versus feedback generation

Despite the technical challenges that remain on the NLP side, there are technical challenges on the side of the feedback generation strategy too. By inspecting the log files we found that often the feedback messages that were generated for a particular learner response reflect that the linguistic analysis contains most of the required information, but the feedback message fails to convey it comprehensibly.

Take, for instance, Figure 11.21, which shows the messages generated for a response to Question 1 in Activity A1-T1. The learner submits the sentence "Salt with water produce Sodium hydroxide,Hydrogen and Clhoride" (sic) and the expected response is something along the lines of "Salt and water produce Sodium hydroxide, Hydrogen and Chlorine."

As shown in Figure 11.21a, corresponding to the feedback at the level of spelling and grammar, the system is able to detect that there is a spelling error and that a punctuation mark is missing. In Figure 11.21b, showing feedback at the level of activity specific language and content, the system is able to detect that the enumeration containing the products of the reaction contains some of the information

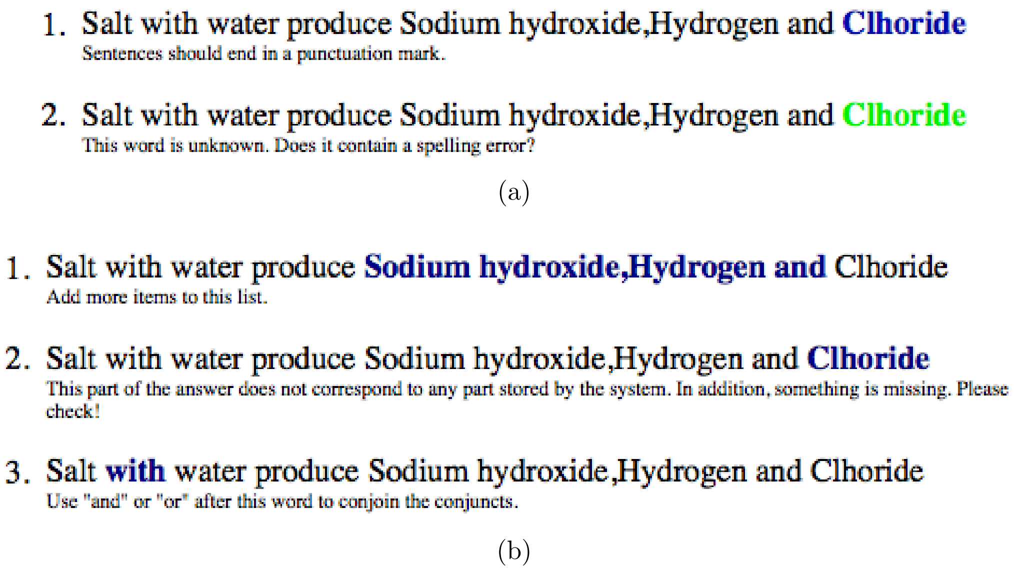


Figure 11.21: Limitations of current feedback strategy Feedback exemplified on a real learner response.

required but not all, and that some of the information in it is incorrect. The figure shows that the reactants should be conjoined using the conjunction “and”, and not the preposition “with”.

However, the way this information is worded into messages for the learner is tricky. Let us take the problems with the enumeration of the reaction’s products. The system provides it in two messages, message no. 1 and message no. 2 in Figure 11.21b. The first message highlights the correct part of the enumeration and reads “Add more items to this list.” The second message highlights the incorrect part of the enumeration and reads “This part of the answer does not correspond to any part stored by the system. In addition, something is missing. Please check!”. The language and information structure used are infelicitous, in technical terms. However, a proper combination of the underlying analysis would allow a message such as “The enumeration contains part of the expected elements, but one of them is incorrect”. More detailed or more explicit feedback messages could be generated. The possibilities here would depend on how much generalised or customisable we would like the strategy to be, but it could certainly be improved.

11.5 Chapter summary

In this chapter, we described an experiment that we carried out to validate and analyse our software concept intended to empower teachers with the methodological and technological instruments to author and employ ICALL activities. The experiment was carried out in a blended learning context with secondary school teachers who are very competent as computer users, and with learners who are used to work with their computers for learning. We also presented the multidisciplinary training that teachers underwent, reflected the experience and the discussions that arose during the process, and presented the results. The results include both the materials created by teachers, a series of learning materials including ICALL materials, as well as the learning experience of learners using these materials.

In a qualitative and quantitative evaluation of the learning experiences, we showed that the NLP-based feedback generation system performs well for the checking of spelling and grammar errors (above 94%), and reasonably well for the checking of task-specific language and content errors (above 60% and up to 76%). We also saw that for some of the feedback messages that were manually validated as correct, learner uptake was observed: Between 40% and 86% of the spelling and grammar error messages had a positive effect, and between 23% and 41% of the activity and language content feedback messages had a positive effect too. As for the percentage of messages with no positive effect, it was below 5% in spelling and grammar errors, it was around 12% in language and content checking for two of the analysed learner groups, and up to 25% for the other group. Moreover, learners did not show any perceptible reaction to a large percentage of the feedback messages, though this cannot always be interpreted as a failure of the system to appropriately assess the learner's response.

In the discussion of the results we analysed teacher and learner satisfaction, and the research point of view. Teachers managed to couple their needs and the identified learner needs to produce pedagogical designs consistent with their course programmes that included automatically corrected activities using an ICALL authoring tool. Moreover, they managed to understand the core of the NLP functionalities provided by the system, and had the impression that the tool helped them conceive activities that allowed them to practice aspects that they could not practice before, or in a way that fostered reflection on the learner side.

Teachers expressed that the material generation process was time and effort consuming, though they admitted that after an initial learning curve the time devoted to the development of ICALL activities is not higher than the time required for other activities, response specification aside. The effort of specifying responses can be offset by recycling activities for several groups and years, as well as by a relatively easy way to enhance the system's correction functionalities. Moreover, the three participant teachers stated that the process as a whole helped them better understand the needs and the behaviour of their learners, as well as the paths their learners follow in order to produced certain linguistic outputs.

As for learners, 53.2% of those who responded to the satisfaction questionnaire expressed that the system's feedback was generally easy to understand and helpful, 38% of them said it was neither easy nor difficult, and 8.8% said it was not easy or helpful. Both teachers and learners agreed that the kind of activity performed resulted in an increase in motivation thanks to the immediate more "intelligent" feedback provided.

Finally, we saw also how the experiment reflected challenges both for the field of NLP and ICALL. The introduction of techniques to further expand the linguistic space abstracted in form of response specifications is probably one of the most outstanding needs in terms of NLP. This can be approached following different strategies. A possible strategy is to use techniques based on the notion of synonymy or meaning similarity to expand the linguistic expressions provided by teachers during the specification process (or during the use of the system during or after the actual instruction). Another possible way is to use techniques based on the notion of inference, such as semantic analysis or rich textual entailment, to expand the human language understanding capabilities of the system.

In terms of ICALL, the feedback generation strategy is certainly an aspect to be improved. Teachers often missed the capability to customise several aspects of the feedback ranging from the possibility to establish the relevance of certain error types on the basis of pedagogical criteria, to the possibility to rephrase canned feedback messages. In terms of feedback generation, we observed how the presence of the appropriate analysis does not always correlate with the generation of a comprehensible message. This finding suggests the functionalities usually attributed to the expert model in Intelligent Tutoring Systems as customisable functionalities should be included in ICALL authoring tools.

Part V

Conclusions

I don't care if the system is not perfect. Most of my pupils used these sessions to think about the task they were working on. Sometimes they even *competed* against the computer [to determine who was right and who was wrong]. To make them reflect on what they do is my primary goal.

February 2012, Teacher 1

If I say that 'I read comics' is a correct answer, then 'I read manga' should be accepted as a correct answer too.

February 2012, Teacher 3

Chapter 12

Conclusions and outlook

In this concluding chapter we discuss the contributions of this thesis with respect to our research goals, as well as the future research lines that we envisage.

12.1 Contributions

As introduced in Chapter 1, we aimed at two goals. The first was to develop a methodology for the design and implementation of ICALL materials taking into account the pedagogical needs and the computational capabilities. Such a methodology should serve as a means to connect the perspective of FLTL, and in particular TBLT, with the perspective of NLP. The second goal was to facilitate the integration of ICALL materials in secondary school instruction settings by developing a technology and the accompanying methodology for teachers to be able to author and use autonomously their own ICALL materials. For each of these goals, we present separate contribution sections.

12.1.1 Connecting TBLT and NLP principles

The first contribution of this thesis is a methodology to guarantee a pedagogically and computationally principled design of ICALL tasks for the development of TBLT materials to be assessed with NLP strategies allowing for the automatic analysis of learner language. This methodology has three principal frameworks: the Task Analysis Framework (TAF), the Response Interpretation Framework (RIF) and the Automatic Analysis Interpretation Framework (AASF).

As the first component in our methodology, the TAF allows for a characterisation of the pedagogical features of FL learning activities from an FLTL perspective. The TAF serves as an initial instrument to specify the degree to which the goals of a FL learning activity, its expected outcome, the envisaged pedagogical and cognitive processes, the desired type of assessment, and its type of response make it compatible with the TBLT approach and a reasonable candidate for the implementation of an NLP-based assessment. It also helps determine the identification of other computer-based assessment strategies using non-NLP techniques – that is, it embraces the characterisation of CALL materials in general.

The second component, the RIF, allows for a detailed characterisation of the linguistic structures that a given ICALL task is expected to elicit from learners. This characterisation includes a description of the relationship between input data and response, as one that provides essential information on the freedom of learners to choose the linguistic resources to complete the task. The RIF crucially includes a formal specification of the thematic and linguistic contents expected in learner responses, as well as a definition of the task's criteria for correctness. As for the characterisation of thematic and linguistic contents, the RIF makes extensive use of notions from descriptive and computational linguistics to describe language. As we suggested, language being the object of study of both disciplines, this seems to be a natural crossroad for NLP and FLTL to meet. Last but not least, the RIF facilitates the characterisation of gold-standard responses that are invaluable for the design and development of NLP-based assessment strategies.

The third component of the methodology proposed is the AASF, which allows for the specification of requirements for the implementation of an NLP-based assessment strategy for a given ICALL task. Our specification process assumes a separation of the language analysis task and the feedback generation task. This separation is crucial for the implementation of modular architectures to develop the functionalities of Intelligent Language Tutoring Systems. Moreover, this modularisation is compatible with the established approaches in ICALL, and facilitates the use of the proposed methodology in more complex ICALL systems, e.g., systems including student modelling modules or more complex expert modules.

In addition to proposing this methodology, a second contribution of the thesis is its practical implementation in a research instruction setting. We described how we applied the proposed methodology to design, develop and implement a multilingual set of CALL materials in the area of business and finance for the learning of foreign languages including NLP-based assessment strategies (we exemplified activities in English and Spanish, but the materials included Catalan and German). The application of the proposed methodology was carried out in a research setting following TBLT as a pedagogical approach, and using finite-state techniques for the implementation of rule-based approaches to the semantic and pragmatic analysis of learner language – where FL learning tasks were understood as a domain of application. We exemplified the implementation of formative and summative assessment on the basis of combining different levels of linguistic analysis using both general and domain-specific NLP tools and resources.

A third contribution within this research goal was the analysis of learner responses to a subset of the materials designed and implemented following our methodology. As we showed, the comparison of elicited learner responses with expected responses for a given task provides linguistic evidence for the analysis of the task's complexity in NLP terms, but also in terms of FLTL. On the basis of such a evidence, we drew further conclusions with respect to the suitability and meaningfulness of the FL learning tasks. As we described, such an analysis can be used to improve and enhance the computational models for the analysis of learner language and the generation of feedback, and it can also help evaluate the pedagogical goals of the learning activity.

A fourth contribution is the further characterisation of the kind of FL learning activities in the so-called viable processing ground. As we discussed, such activities includes tasks with different levels of complexity, as well as different types of assessment required. Moreover, these activities present different characteristics in terms of variation, both at the level of contents and the level of form, which seem to have a correlation with the relationship between input data and response, concepts that can be used to inform FLTL and SLA of the pedagogical characteristics of the tasks that are suitable for NLP-based automatic assessment.

Finally, a fifth contribution is the exemplification of how the process of designing, implementing and employing of FL learning materials can be conceived as a cyclic approach to the development of ICALL tasks. Through this approach, the pedagogical and the computational requirements of the task can be incrementally and iteratively improved and refined, favouring its re-usability and recycling possibilities. Interestingly, this approach to the creation of ICALL materials is in line with research and practice in the fields of FLTL and CALL (Estaire and Zanón, 1994, Willis, 1996, Colpaert, 2006).

12.1.2 NLP as an enabling technology for teachers

As for the second goal, the integration of ICALL materials in instruction settings, the first contribution of this thesis is the methodology and the technology proposed through which teachers can autonomously design, implement and use in class FL learning activities including NLP-based automatic assessment functionalities. This strategy includes three elements: (i) the Response Specification Language (RSL), (ii) a strategy for the expansion of correct well-formed responses to a given activity to make the ICALL system capable of handling a range of varying, or deviating, responses, and (iii) the Response Specification Scheme (ReSS).

As for the RSL, it is a formal language through which the expected responses to a given FL learning activity can be specified in a form and structure containing the minimal information needed to automatically generate the NLP resources required for the customisation of NLP-based assessment functionalities. The RSL is a non-metalinguistic interface between the specification needs of NLP-based feedback generation architectures and the establishment of criteria for correctness under a pedagogical perspective. It is therefore another stone in the building of the bridge to connect FLTL and NLP, in this case in practice.

The second component of this methodology is the automatic expansion of correct well-formed responses into a range of linguistic models to handle a variety of deviating structures to be included in a customisable approach to feedback generation. The expansion based on the teacher-specified RSL-formatted responses results from applying that standard linguistic change and surface transformation operations which are well-known and common in the characterisation of learner language (Corder, 1981, James, 1998). The expansion techniques can therefore be sensitive to corpus-based findings in FLTL and SLA studies.

The third component of this methodology is the Response Specification Scheme, a methodology that makes it possible for teachers to specify activity responses in RSL format on the basis of regular human capabilities, as opposed to programming code as a learnt capability. By exploiting the notions of paradigmatic and syntagmatic relations, and assuming the corresponding graphical interface (commented on below), the ReSS allows teachers to organise responses in a way that they result into RSL-valid declarations.

Complementing these conceptual component, the second contribution to the integration of ICALL materials in class is the evaluation of the authoring tool and the methodology we proposed in secondary school instruction settings. The experiment supposed an active collaboration with teachers and learners in their own instruction settings, which were characterised by the use of a blended learning approach with support of a particular learning management system The teachers and the students that participated in the experiment had a reasonable expertise in the use of computers, and were generally motivated to use technology for educational purposes. This particular experiment setting allowed us to take into account the perspective of the teacher and the learner throughout this research.

From this experiment, we were able to determine that teachers are capable of generating ICALL materials including NLP-based automatic assessment with a reasonable amount of effort and within a reasonable amount of time – and with the perspective of an iterative improvement of the activities through sustained use. Moreover, the resulting materials can be integrated in the language programme, despite the finding that teachers might prefer to better know the behaviour of the ICALL system before they fully integrate it in the class's workflow. Last but not least, we showed that teachers were able to transform their working methodologies by means of reflection on the available technologies and on the adaptation of such technologies to the needs of their learners.

An important feature of the experiment is that we took into account the teacher and learner subjective view of the experience. As a result, we could see how teacher expectations change over time, as well as how learners can critically distinguish between aspects of their learning that are being supported by the proposed methodological innovations. In general, the experiment showed that both teachers and learners were motivated by a context in which they believed to have, and they actually had, a certain room for manoeuvre and opinion – a finding in line with Levy (1997: p. 97)'s argument in favour of teacher control as a guarantee for target learner appropriateness of the materials and for the motivation of learners.

In this respect, the research we presented qualifies to some extent as *action research* (Nunan, 1992: pp. 17–19), a research approach in language learning in which university-based researchers and FLTL practitioners collaborate with the goal to improve and change certain aspects of the instruction setting. As a matter of fact, we firmly believe that the procedures proposed in action research, which critically follow again a cyclical approach from problem design to evaluation and dissemination, can be a very fruitful path to follow among CALL and ICALL researchers. As a collaborative research-practice approach, it can strengthen the applied and the theoretical side of our work.

Finally, we showed that, generally speaking, the feedback generation system performed reasonably well, though in the formal aspects of language it performed better than in the thematic aspects of language. The analysis of the system's feedback shows that state-of-the-art off-the-shelf NLP software can provide useful assessment functionalities, which are, to a certain extent, comparable with the feedback that a teacher would provide learners with.

12.1.3 General contributions

This thesis contributes at a more general level to the three research areas on which it focuses: ICALL, FLTL and NLP.

In terms of ICALL, this thesis supposes a sound step toward the conceptualisation and the theoretical underpinnings of the design and development of ICALL materials and the different aspects to be taken into account in the larger multidisciplinary context of ICALL. Our theoretical and methodological proposals are accompanied by two practical empirical studies, through which the relevance of data becomes even more apparent, both as a key to understanding the dynamics of learner reactions to activity instructions and context and as a key to improve and enhance the strategies for the automatic analysis of learner language.

In terms of FLTL and NLP, this research further supports the need to conjoin efforts in areas such as ICALL to be able to put into practice theoretical and practical principles in real-world human activities. Moreover, our research shows that much is to be gained from involving all the agents in the teaching/learning setting. As a procedure, our research approach genuinely turns real-life problems as one of the motors of applied research. Finally, we suggested new challenges and research for avenue for FLTL and NLP which, independent of ICALL, already have an interest in the respective fields.

12.2 Future work

There are two different types of future work that we envisage. On the one side, we envisage a series of research lines to improve the different research methodologies and the feedback generation strategies that we propose in this thesis. On the other side, we foresee two interesting longer term research goals to be pursued in the field of ICALL.

12.2.1 Thesis-related short term research

In terms of the characterisation of the viable processing ground, a logical next step would be to develop quantitative-qualitative measures to assess the pedagogical complexity and the computational feasibility of ICALL tasks. Following the exploration of learner responses presented in Chapter 9, linguistic features of the expected and the elicited responses could be used to help FLTL and SLA researchers assess the complexity of the tasks, and to support NLP researchers in assessing the complexity and the characteristics of the language processing strategies to be followed. In this respect, a particularly interesting line of research would be the collaboration with SLA researchers investigating on the effects of task complexity on the learning of second languages – see (Robinson, 2011).

As for the NLP-enhanced methodology for the authoring of ICALL materials, a series of practical improvements could be implemented. First, the design-based response specification process should simplified as much as possible, and made compatible with an incremental enhancement by profiting from individual learning experiences. A way of simplifying the process is to introduce NLP-rich techniques to facilitate the search for expressions with semantic similarity. Thus, provided a verb or a verb phrase, the system could search for synonymous expressions via dictionaries and domain-specific corpus. A strategy to profit from learner responses to enhance teacher specifications and feedback generation would require teacher functionalities to benefit from ongoing collected responses. With the adequate functionalities teachers could use the time devoted to review learner responses to a given activity to easily increase the spectrum of correct responses, or the fine-grainedness of particular feedback messages.

Though in its actual development the ReSS-based response expansion process is a way of empowering teachers to author ICALL activities, this expansion processes could be improved. This is a task that should be done on the basis of corpus-based research, so that the different interlanguage levels and learner profiles can be characterised. This research would necessarily be carried out in collaboration with researchers in Second Language Acquisition.

A third aspect to work on is the refining of the assessment functionalities of the ICALL material authoring tool. The presentation of feedback to learners should be further customisable according to teacher criteria: This will include the customisation of the graphical presentation of feedback, but also of the actual wording of the feedback. A particular interesting line to investigate would be to connect the thematic contents of teacher specifications with feedback messages assessing the meaning, not only the form, of the response. This could be pursued by using the metainformation associated with the Response Components as part of the feedback messages to be generated.

Fourth, an interesting line to continue with would be the comparison of system assessment versus teacher assessment. Such studies would facilitate the evaluation of the NLP-based feedback generation software, but they would also promote the transferring of teacher practices with respect to correction to the assessment module.

12.2.2 Longer term research in ICALL

This thesis suggests two long-term research lines to be pursued in the future. One of them should define a more comprehensive and detailed methodology for the development and analysis of CALL materials, and in particular one that links the perspectives of the different disciplines involved with the perspective of the pedagogical and linguistic goals of FL learning tasks. Such a methodology would have to be flexible enough to include the features that concern the development of ICALL materials as the TAF, the RIF and the AASF do – mainly pedagogical and computational features in terms of language. However, it should also include room for the characterisation of learner profiles and styles, the characterisation of teaching strategies, maybe even adaptive teaching strategies, and the establishment of assessment procedures. Crucially, incorporating these dimensions to a material development methodology should not be specifically made for ICALL, but be compatible.

In this respect, a research direction to explore is the one suggested in Colpaert (2006) for CALL in general, which has already been suggested as an interesting line to follow in Schulze (2008). According to these authors, CALL and ICALL materials are better integrated in instruction settings if (i) the inclusion of technology is taken into account from the beginning and (ii) if the process is iteratively and cyclically evaluated and improved. Such a research direction would be in line with proposed methodologies in the design of task-based instruction materials (Estaire and Zanón, 1994; Willis, 1996), with the incorporation of corpus-based decisions and observations as part of the ICALL material's life cycle, and the enhancement of NLP-based response analysis and assessment strategies based on empirical evidence.

The second research line would pursue to make practical the use of NLP-based strategies for the customisation of automatic assessment functionalities. Our findings show that teachers are both capable and eager to profit from technologies using Artificial Intelligence to conceive new methodologies that help them and their learners achieve a greater autonomy in the teaching/learning task. However, our findings also reflect a major need to improve essential parts of the process and the methodology we propose for the teacher-driven generation of NLP resources. Teachers profit from the capability to tailor feedback generation functionalities, but they must be able to determine the type, the appearance, and the wording of feedback messages. This suggests the need to evolve from the Response Specification Language to something we want to call the Assessment Specification Language. In other words, there is a need for teachers to be able to control and interact with higher level computer functionalities, the need for teachers to have a means to operate with computers – like a control panel. This is a research line that would require very close interaction with teachers, and one that would neatly fit in the so-called action research programme.

Appendixes

Appendix A

ALLES learning units: final tasks and task sequencing

This appendix includes the detailed description produced for two of the learning units developed during the ALLES project following the initial steps of Estaire and Zanón (1994)'s framework for the design to task-based instruction materials. The two units correspond to the B2 and C1 level units on the topic *Career Management and Human Resources*. The respective titles are *Education and Training* and *Job Interview*. The description of the all the units developed during the project can be found in (Díaz, Ruggia, and Quixal, 2003a).

1 Career Management and Human Resources

1.1 Education and Training (B2)

1.1.1 Final Task

At the end of the unit, the student will **write** an email where he will register for a training course offered at his company. In this email the student will specify reasons why he is interested in taking this course and the timetable.

To complete this task, the student will use:

1. The course listing attached by Human Resources to the email describing the availability of training courses
2. His schedule for the current month
3. Voice mail from his boss recommending a particular course

1.1.2 Unit objectives

During the unit the students will develop, with a degree of communicative competence in accordance with their level, the ability and knowledge necessary to:

- Understand requirements to register for courses.
- Write emails in order to complete a registration.
- Speak about her or his interest.
- Know how to write professional emails (structure, expressions, tone, etc.).

1.1.3 Contents necessary to carry out the final task

Thematic content

- Registering for courses
- Professional emails

Linguistic content

- Lexical: words, expressions and gambits used for registration, courses and schedules.
- Functional content: expressing likes and dislikes, making suggestions, writing an email (techniques, structure, control…), recommending and asking for advice, describing (courses).

- Grammar content: grammar structures used for making suggestions, recommendations, asking for advice, describing things.
- Textual types: registration forms and e-mails

Socio-cultural

It will fit the material collected for this unit

1.1.4 Process plan

Subtask 1 (main skill: reading)

The student will **read** a business article regarding the importance of having a properly trained workforce and value of human capital in the companies. Next, the student will **read** various work schedules from different employees in a company, their job profiles and a list of specialised courses offered by the Human Resources department. They have to match the employees' schedules and profiles with the courses they could take for further advancement in their careers and explaining why these matches are appropriate.

Subtask 2 (main skill: writing; other skills: listening, reading)

The student will listen to a recording of an informal talk between two employees exchanging views on different training courses offered at their company and discussing pros and cons. Next, the student will read some short articles on the use of emails in business settings and how to write formal and informal emails. Finally, the student will **write** a short informal email to a friend. The email topic will be a description of courses listed on a leaflet and questions about what courses to take.

Subtask 3 (main skill: speaking)

The student will do a **role-play** activity in which they will call human resources department asking for seat availability for a particular course, use of laptop during the course, material required, and whether there will be a diploma issued at the end.

1.2 Job Interview (C1)

1.2.1 Final task

At the end of the unit, the student will **have a job interview** with the Human Resources Manager of a company.

To complete this task, the student will use:

1. A job announcement
2. His CV
3. A letter of presentation.

1.2.2 Unit objectives

During the unit the students will develop, with a degree of communicative competence in accordance with their level, the ability and knowledge necessary to:

- Use their active skills: The student writes her/his own CV.
- Be able to give information about everything that is relevant for job application such as training, schools, universities, qualifications, professional life, personal life (hobbies).
- Be able to present himself or herself orally in a favourable way.
- Be able to respond in conversations about job responsibilities etc.
- Be able to participate in a conversation with the right register.

1.2.3 Contents necessary to carry out the final task

Thematic content

- School systems, professional training, university training in a certain country
- Job responsibilities, qualifications, careers
- Communication: how to present her-/himself?
- Appropriate communication in a job interview. Analyse and answer the questions of an employer. Appropriate communication, registers and communication strategies.

Linguistic content

- Lexical content (vocabulary related to the school system and CV, training, career)
- Functional content: understanding job ads, self presentation, descriptive abilities concerning training, career.
- Textual types: CV, presentation letter

Socio-cultural content

- Different school systems and denominations of degrees, interview situations.

1.2.4 Process plan

Subtask 1 (main skill: listening; other skills: writing)

The student will **listen** to five job ads from different companies. (Areas: Computer scientists, translators, sales persons, receptionists, managers, craftsman).

Listening comprehension exercises.

After this, the student will write a scheme taking into account the most relevant information in each case.

Subtask 2 (main skill: reading)

The student will read some articles about facts to be considered by an applicant during a job interview. He will write his own studies and his professional experience in order to prepare an eventual CV.

Subtask 3 (main skill: speaking; other skills: listening, reading)

The student will listen to a job applicant reading his CV. Then, the student will read some recommendations about drafting a CV and afterwards he will record his own CV according to these recommendations.

Subtask 4 (main skill: writing; other skills: reading)

The student will read some articles about how to prepare a presentation letter (Reading comprehension exercises). After this, he will **write** a presentation letter (Writing exercise)

Subtask 5 (main skill: speaking; other skills: listening)

The student will listen to a job interview with an employer. The student has to analyse the reaction of the applicant and record his remarks. Then, the student will record a voice mail explaining how well the applicant did during the job interview.

Appendix B

On finite state machines

A finite-state machine (FSM), or finite-state automaton (FSA), is a mathematical abstraction often used to design computer programs. It is a behaviour model composed of a finite number of states, transitions between those states, and actions, which allows to model a flow graph in which one can inspect and monitor the way an agent proceeeds when certain conditions are met. Finite-state machines provide a simple computational model with many applications.

Formally speaking a (deterministic) finite state machine consists of the following:

- Σ a finite set of symbols, known as the *input alphabet*;[1]
- Q, a finite set of *states*;
- $i \in Q$, a particular state called the *initial state*;
- $F \subseteq Q$, a set of *final states* ;
- $\delta : Q \times \Sigma \to Q$, a function δ from $Q \times \Sigma$ to Q, called the *transition function*.

Usually, the machine starts in the initial state i. The input is a string of characters from the input alphabet which are read one at a time (from left to right). At each stage the machine is in some state $s \in S$. If the machine is in state s_i, and the next input character is $c \in I$, the machine moves to state s_j as a result from applying the function $\delta(s, c)$ and awaits the next input character. The process continues in this way until all the input characters have been processed.

A specific type of FSM are the so-called Finite State Transducers. *Transducer* is a term opposed to *acceptor*, which is the type of FSM we just described. The main characteristic of transducers is that in addition to process the sequence of symbols according to the transition function (δ), they are capable of generating an output once they have reached a (final) state. Once the final state has been reached transducers may produce more than one output – and then they are non-deterministic transducers – or they can give no output at all.

[1]Note that the use of terms *input* and *output* as used in this section have nothing to do with the meaning that they have as terms in FLTL.

Finite state transducers have been extensively used in computational linguistics because, in general, they compute a relation between two formal languages. Natural Language Processing tools basically parse (process) a set of symbols (natural language) in order to generate a linguistic analysis (a formal language in itself).

Appendix C

ALLES materials as presented to learners

In this appendix we include a reproduction (through screenshots or typed from scratch) of those audiovisual and textual materials learners are exposed to in the ALLES learning tasks mentioned in in Section 7.2 in Chapter 7 of this thesis.

C.1 Screen captures of *Stanley Broadband customer satisfaction questionnaire*

These are the screen captures for the task described in Section 7.2.2.1 as accessed by learners in the ALLES site.

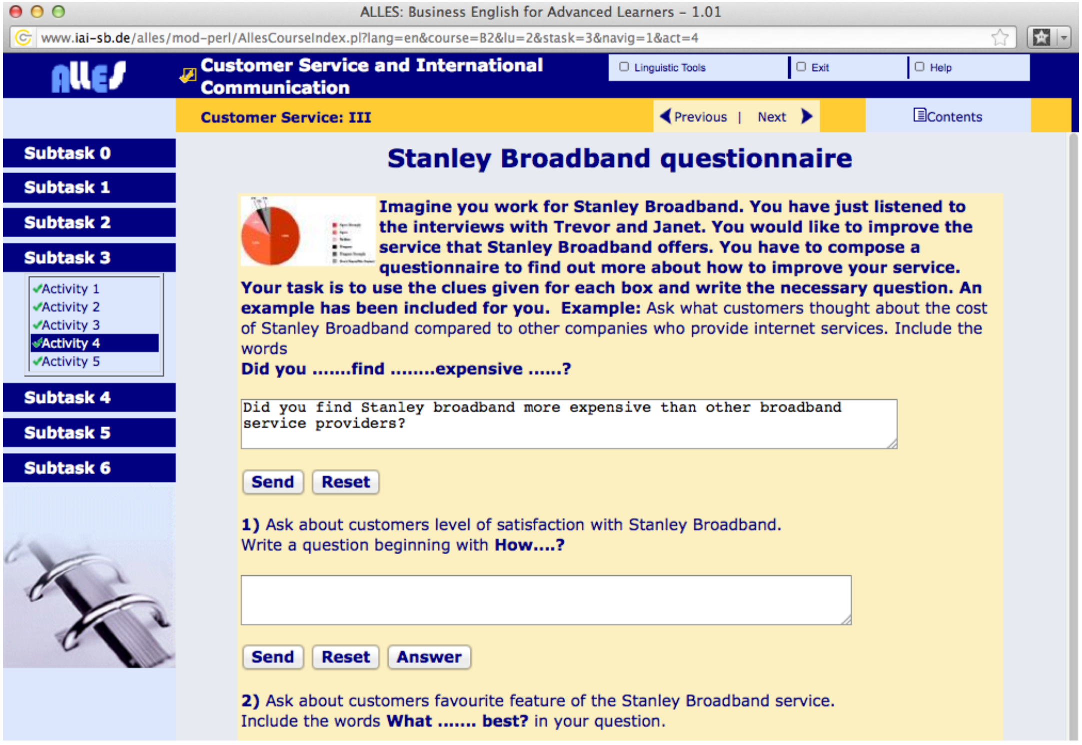


Figure C.1: Screen capture of the overview of the task “Stanley Broadband customer satisfaction questionnaire”.

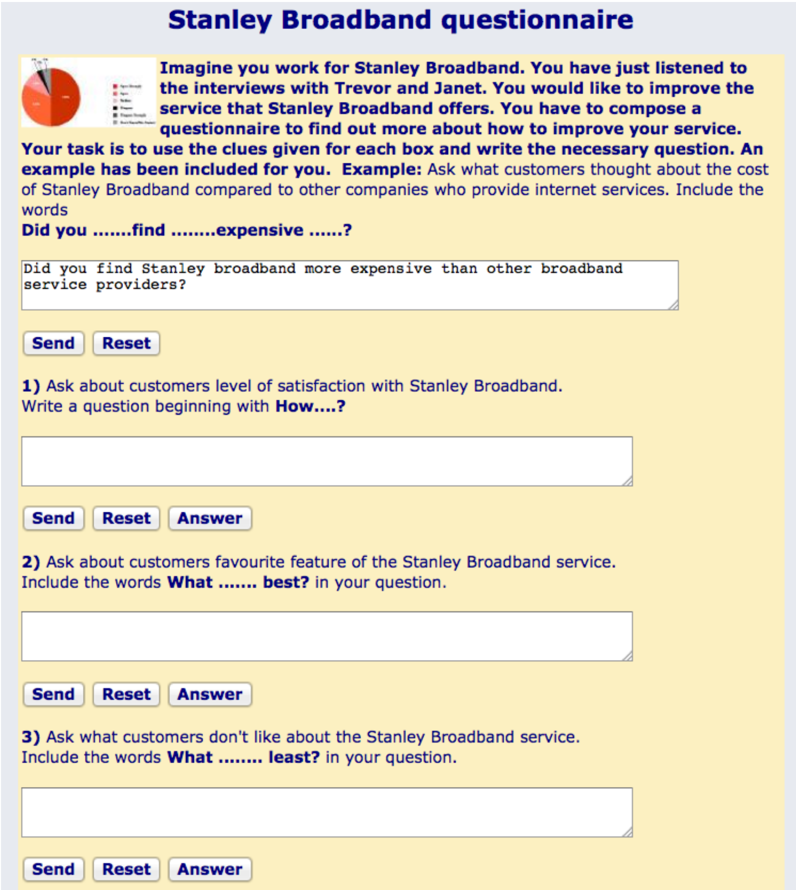


Figure C.2: Screen capture of the details of the task “Stanley Broadband customer satisfaction questionnaire” (I).

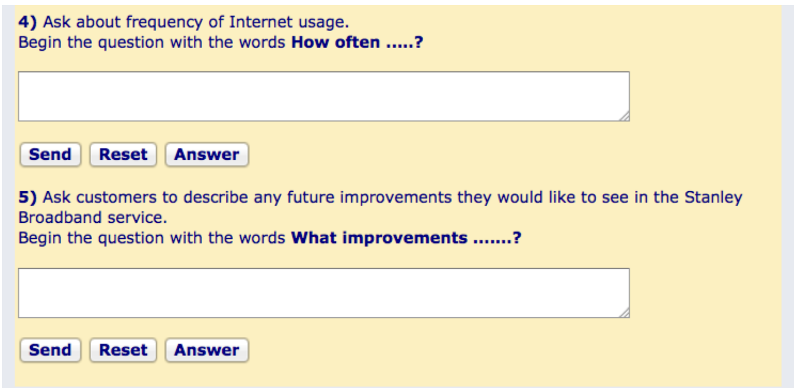


Figure C.3: Screen capture of the details of the task “Stanley Broadband customer satisfaction questionnaire” (II).

C.2 Screenschots of *Describe the structure of your company to a colleague of yours*

These are the screenshots for the task described in Section 7.2.2.2 as accessed by learners in the ALLES site.

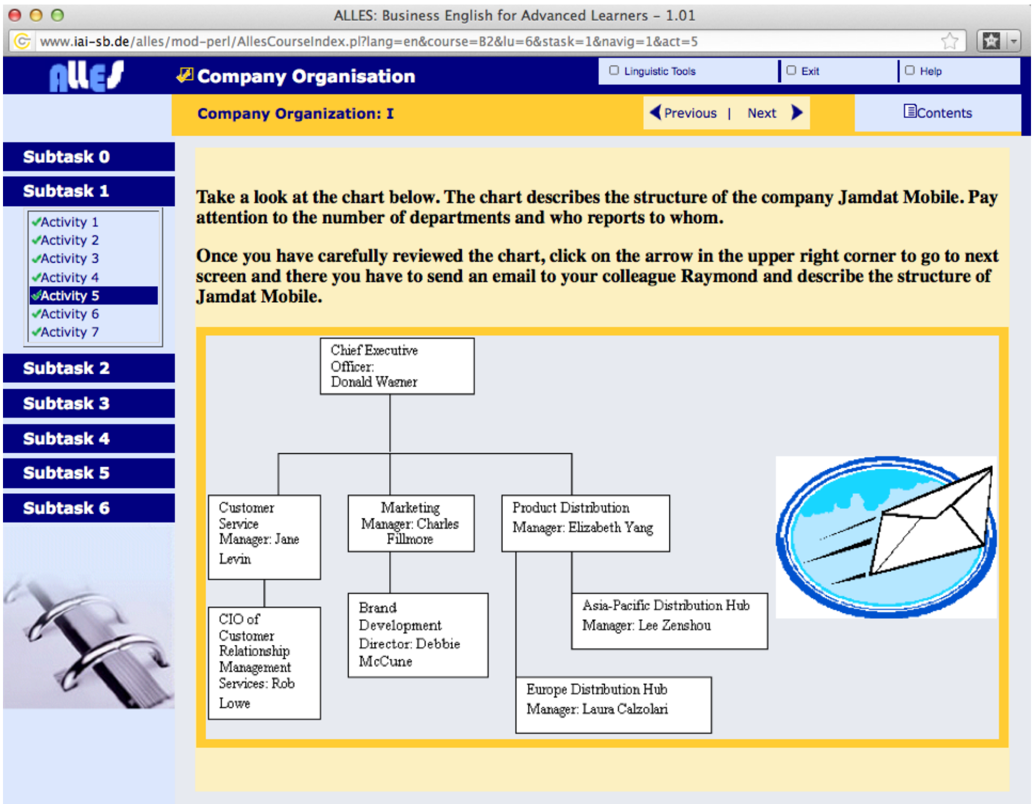


Figure C.4: Screen capture of the overview of the task "Describe the structure of your company to a colleague of yours", Activity no. 5.

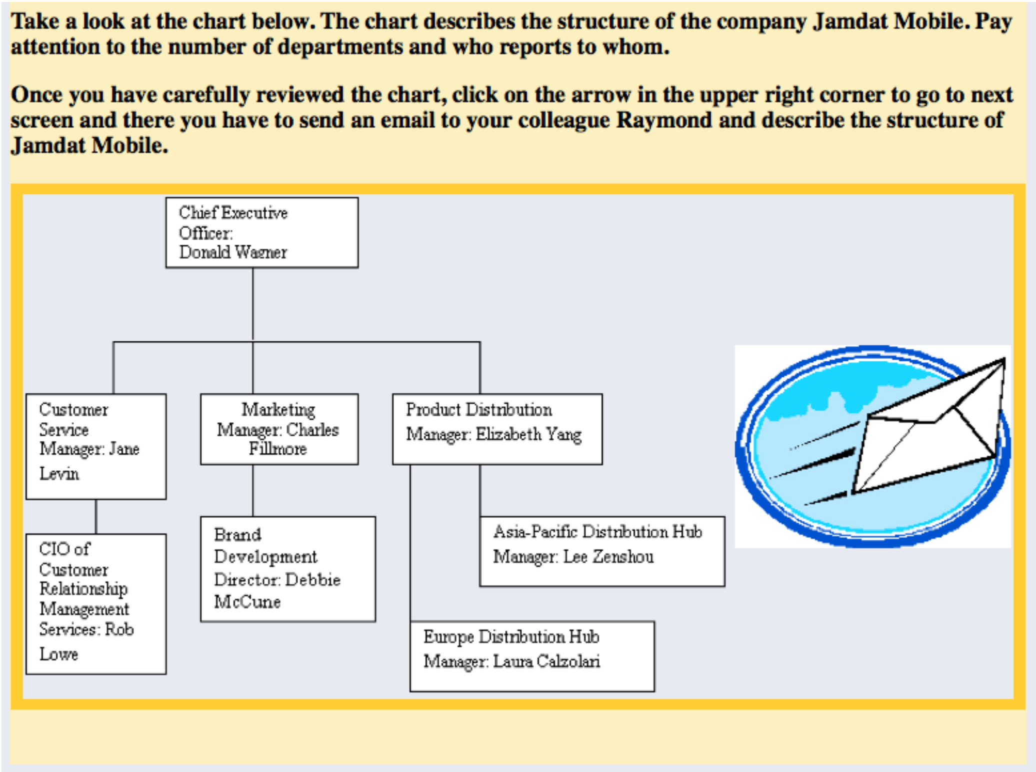


Figure C.5: Screen capture of the details of the task "Describe the structure of your company to a colleague of yours", Activity no. 5.

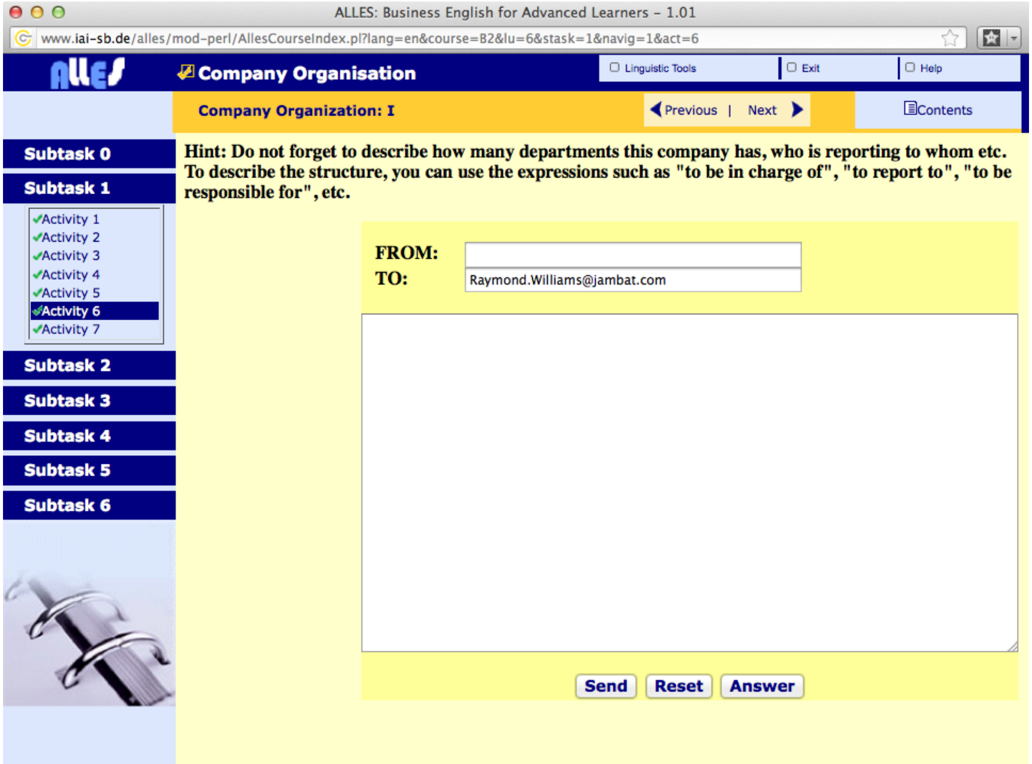


Figure C.6: Screen capture of the overview of the task "Describe the structure of your company to a colleague of yours", Activity no. 6.

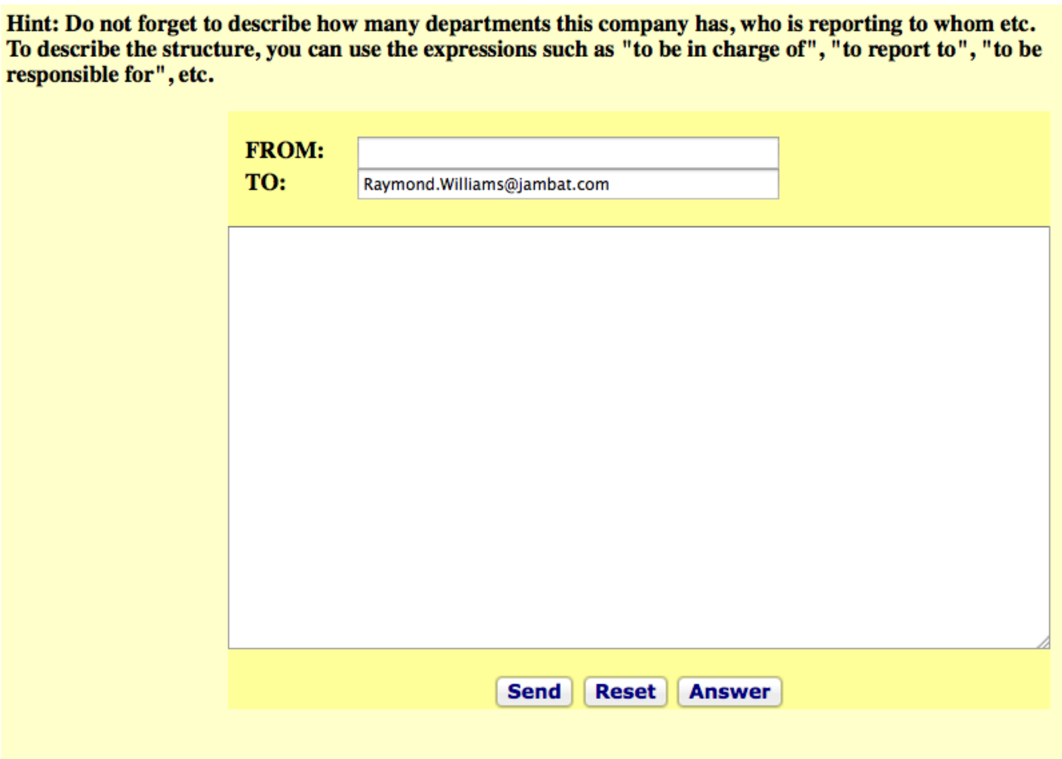


Figure C.7: Screen capture of the details of the task "Describe the structure of your company to a colleague of yours", Activity no. 6.

C.3 Screenshots of *Registering for a course*

These are the screenshots for the task described in Section 7.2.2.3 as accessed by learners in the ALLES site.

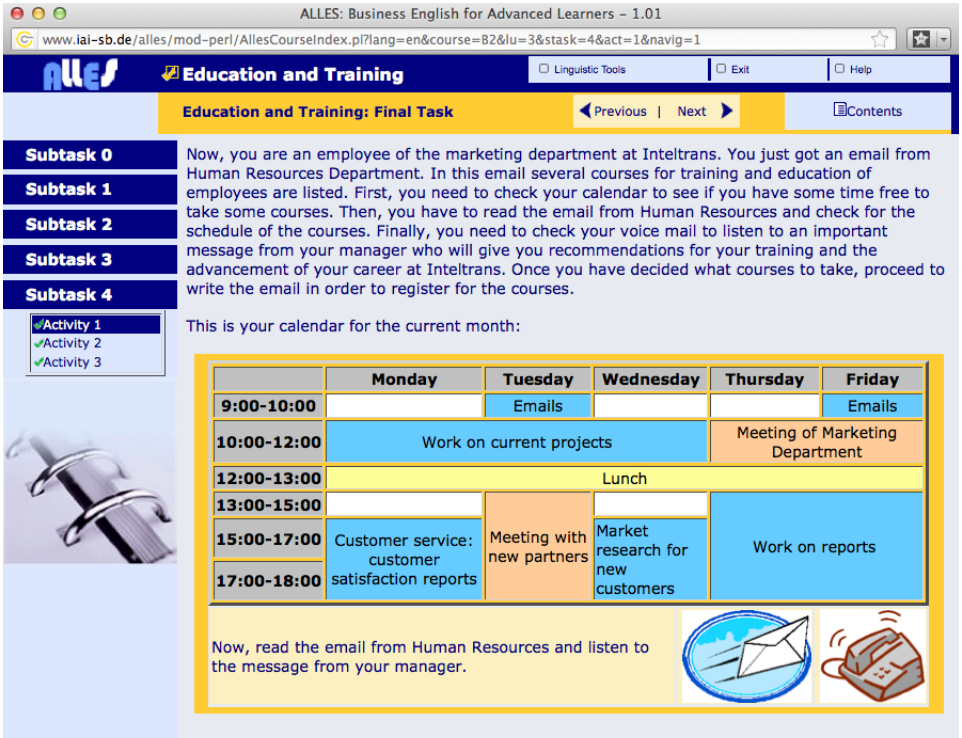


Figure C.8: Screen capture of the overview of the task “Registering for a course”, Activity no. 1.

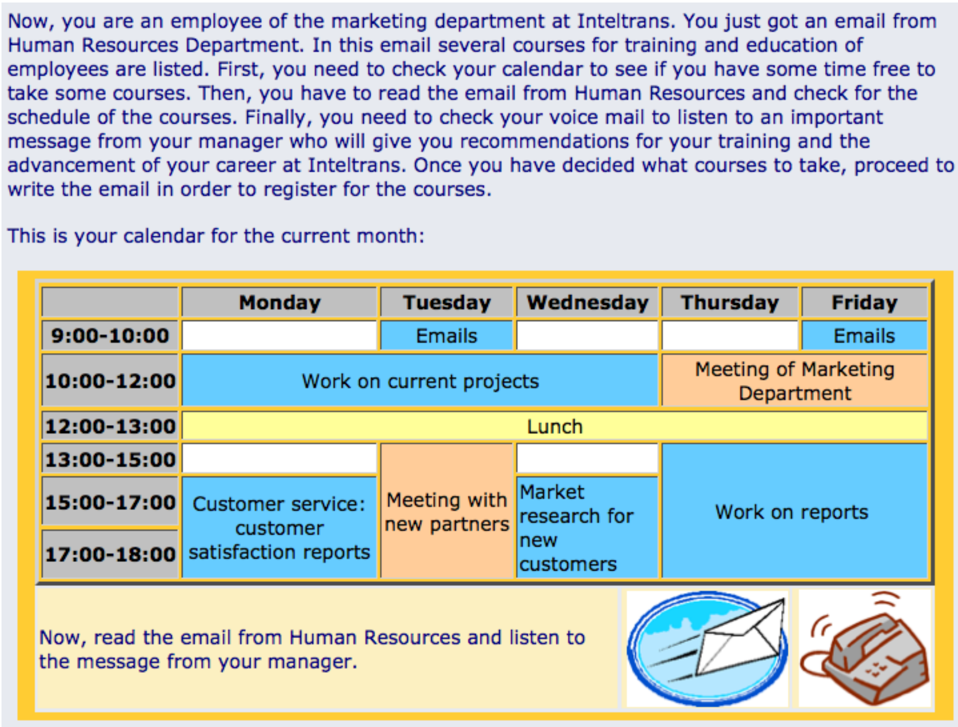


Figure C.9: Screen capture of the details of the task "Registering for a course", Activity no. 1.

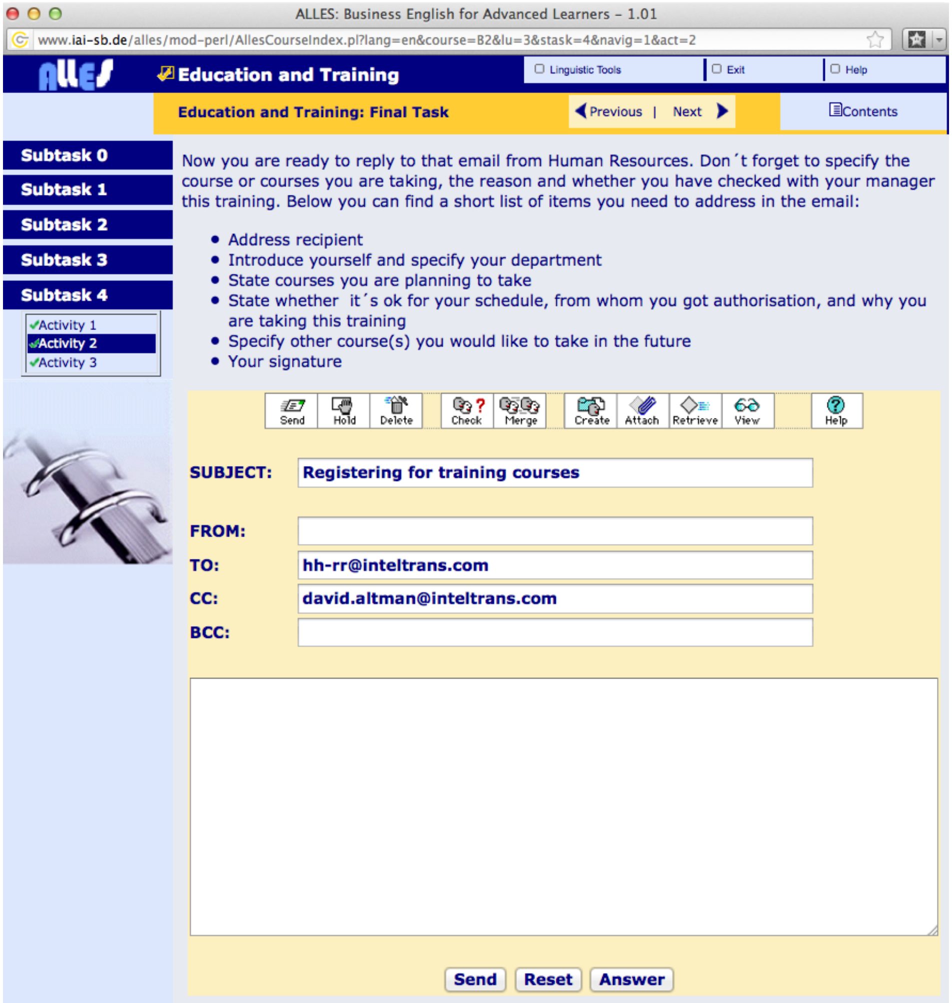


Figure C.10: Screen capture of the overview of the task “Registering for a course”, Activity no. 2.

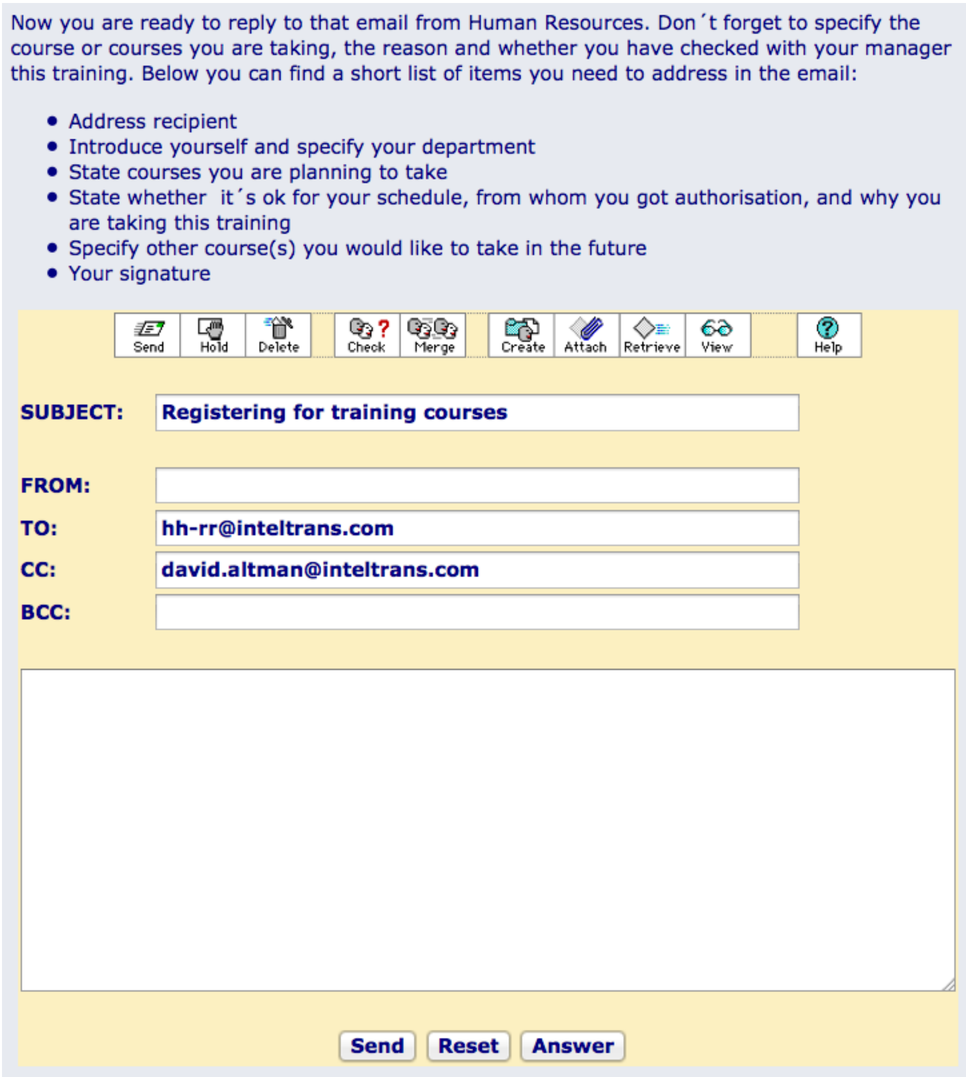
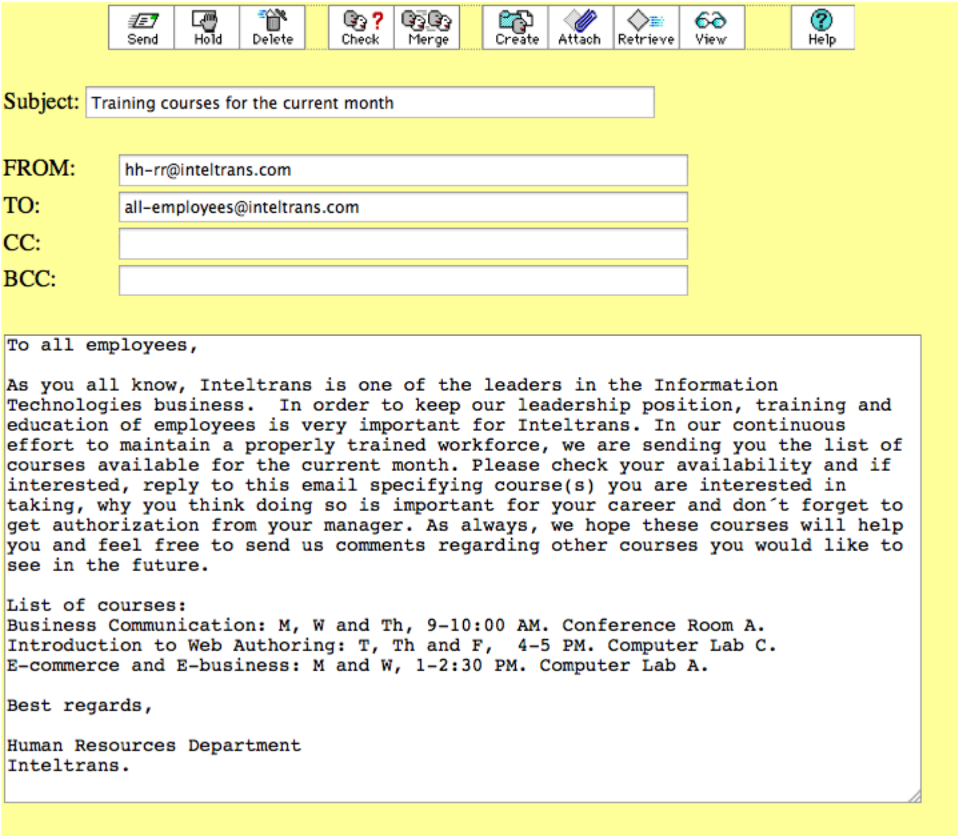


Figure C.11: Screen capture of the details of the task "Registering for a course", Activity no. 2.

C.3.1 Input data included in the activity

C.3.1.1 "Email from the Human Resources Department'


Send Hold Delete Check Merge Create Attach Retrieve View Help

Subject: Training courses for the current month

FROM: hh-rr@inteltrans.com
TO: all-employees@inteltrans.com
CC:
BCC:

To all employees,

As you all know, Inteltrans is one of the leaders in the Information Technologies business. In order to keep our leadership position, training and education of employees is very important for Inteltrans. In our continuous effort to maintain a properly trained workforce, we are sending you the list of courses available for the current month. Please check your availability and if interested, reply to this email specifying course(s) you are interested in taking, why you think doing so is important for your career and don´t forget to get authorization from your manager. As always, we hope these courses will help you and feel free to send us comments regarding other courses you would like to see in the future.

List of courses:
Business Communication: M, W and Th, 9-10:00 AM. Conference Room A.
Introduction to Web Authoring: T, Th and F, 4-5 PM. Computer Lab C.
E-commerce and E-business: M and W, 1-2:30 PM. Computer Lab A.

Best regards,

Human Resources Department
Inteltrans.

Figure C.12: Screen capture of the email given as input data to the learner in Task "Registering for a course".

C.3.1.2 "Message from your manager"

Transcription of the message that learners actually had to listen to:

> Hi, this is David Altman, your new manager. I've been reading your curriculum, and since the Human Resources Department is offering some interesting courses on Information Technologies this month, I recommend that you have a look at the Business Communication and E-Commerce courses. I think they could be useful for the marketing projects we will have to develop by the end of the year. Take a look at your calendar and copy me in the email when you write back to Human Resources. That's all! Thank you! Hmmm... one more thing. I was very impressed at your presentation today. We'll talk some more about it when I get back. Congratulations! I'll see you in a few weeks!. Bye!

C.4 Screenshots of *Expresa tu satisfacción o insatisfacción con el producto* Smint

These are the screenshots for the task described in Section 7.2.2.4 as accessed by learners in the ALLES site.

Figure C.13: Screen capture of the overview of the task "Expresa tu satisfacción o insatisfacción con el producto *Smint*".

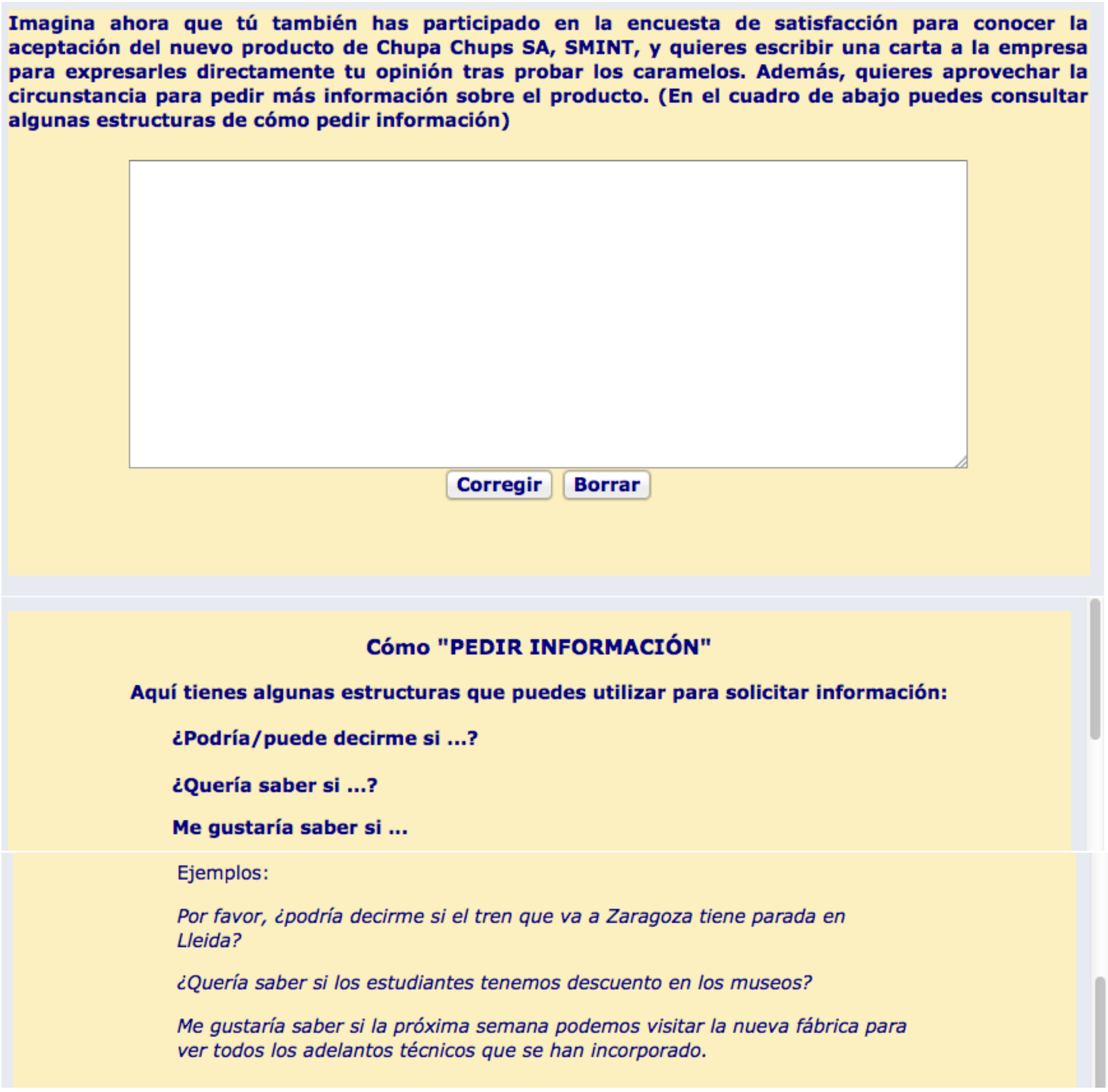


Figure C.14: Screen capture of the details of the task "Expresa tu satisfacción o insatisfacción con el producto *Smint*" (I).

Appendix D

Lexical measures for the assessment of specific vocabulary

As we said in Section 8.5.1, in ALLES, the evaluation of the indicators obtained from a learner response for which summative assessment was required was to be performed against reference values. Ideally, these reference values should be obtained from statistically studies based on large corpora obtained from learners responding to the same activities in the same level and similar learning circumstances. The complexity of such a procedure excluded the possibility to do it so during the life of the ALLES project, and was substituted by indicators defined by content designers on the basis of their expertise.

The only indicators for which a more experimental approach has been trialled is the evaluation of the use of the specific vocabulary, a sub-part of the dimension lexical contents. The strategy used to do so is inspired in the strategies used in document retrieval tasks. The technique basically consists in using a simple statistic to measure how salient certain words in a text are compared to all the words in the text (or in a collection of texts).

The system considers a text as a vector representation. Each vector component is a lemma and the vector's dimension is as big as the number of lemmata that have been identified as relevant for that activity. Each lemma in the vector is assigned a relevance value according to a list, which is activity-specific. The relevance values were manually assigned by content designers, but ideally a corpus-based strategy should be used to determine which lemmata are more relevant than others and to what extent. With this we generated a weighed vector representation of the lexical contents of the response.

To measure how close the response vector is to a reference vector. The reference vector results from combining the vectors obtained from three texts provided by expert writers, non-native speakers of English with a level higher than B2 in the CEF.

The formula used to compute the distance between the response vector and the reference vector is reflected in Equation (D.1). Given a vector $\vec{R}$, a reference lexical representation in form of a vector, and a vector $\vec{L}$, which is the lexical representation of a learner response, the distance between them is the sum of the vector component multiplication divided by the product of the square roots the sums of the component square of each vector. The value of this distance ranges from 0 to 1; the closer to 1, the most similar, and, therefore, the better.

$$sim(\vec{R}, \vec{L}) = \frac{\sum_i (r_i \cdot l_i)}{\sqrt{\sum_i (r_i)^2} \cdot \sqrt{\sum_i (l_i)^2}} \tag{D.1}$$

In ALLES values under 0.6 are considered low enough to judge the learners' text inadequate in this particular feature. Values between 0.6 and 0.8 are considered acceptable, and values above 0.8 are considered similar to native-speaker production. These values are correlated with the percentages required by content designers in the assessment tables (see Table 7.13).

Appendix E

Detailed NLP specifications for activities of Type I, III and IV

E.1 NLP specifications for Task *Customer Satisfaction and International Communication*

E.1.1 Specified correct well-formed responses for Item 1

(76) a. How happy/satisfied are/were you with the Stanley Broadband service?
b. How happy/satisfied are/were you with Stanley Broadband?

E.1.2 Variations on specified responses for Item 1

1. Omission of determiner or preposition
2. Substitution of determiner by another determiner
3. Substitution of preposition by another preposition
4. Omission of *service*, when *the* is present
5. Omission of interrogation mark
6. Omission of capital letter in *How*
7. Substitution of *happy/satisfied* by a different adjective
8. Wrong order for subject and predicate

E.1.3 Specified correct well-formed responses for Item 2

(77) a. What do/did you like least about the Stanley Broadband service?
b. What feature of the Stanley Broadband service do/did you like least?
c. What do/did you like least about Stanley Broadband?
d. What feature of Stanley Broadband do/did you like least?

E.1.4 Variations on specified responses for Item 2

1. Omission of determiner or preposition
2. Substitution of determiner by another determiner
3. Substitution of preposition by another preposition
4. Omission of *service*, when *the* is present
5. Omission of interrogation mark
6. Omission of capital letter in *What*
7. Wrong order for subject and predicate

E.1.5 Specified correct well-formed responses for Item 3

(78) a. What do you think is the best feature of the Stanley Broadband service?
b. What feature of the Stanley Broadband service do/did you like best?
c. What do you think is the best feature of Stanley Broadband?
d. What feature of Stanley Broadband do/did you like best?

E.1.6 Variations on specified responses for Item 3

1. Omission of determiner
2. Substitution of determiner by another determiner
3. Substitution of preposition by another preposition
4. Omission of *service*, when *the* is present
5. Omission of interrogation mark
6. Omission of capital letter in *What*
7. Wrong order for subject and predicate
8. Use of Saxon genitive with feature, instead of "of"

E.1.7 Specified correct well-formed responses for Item 4

(79) a. How often do you use the Internet?

E.1.8 Variations on specified responses for Item 4

1. Omission of determiner
2. Substitution of determiner by another determiner
3. Omission of *Internet* or *often*
4. Omission of interrogation mark
5. Omission of capital letter in *How*
6. Wrong order for subject and predicate

E.1.9 Specified correct well-formed responses for Item 5

(80) a. What improvements would you like to see in the Stanley Broadband service in the future?
b. What improvements would make the Stanley Broadband service better?
c. What improvements would you like to see in Stanley Broadband in the future?
d. What improvements would make Stanley Broadband better?

E.1.10 Variations on specified responses for Item 5

1. Omission of determiner *the* or preposition *in* (both occurrences)
2. Substitution of determiner by a different one
3. Substitution of preposition by a different one
4. Omission of *service*, when *the* is present
5. Omission of *improvements*
6. Omission of interrogation mark
7. Omission of capital letter in *What*
8. Substitution of *make* or *see* by other verbs
9. Wrong order for subject and predicate

E.2 NLP specifications for Task *Registering for a course*

E.2.1 Specified correct well-formed versions of component "Greeting"

(81) a. To Human Resources:
b. Dear Sir(s), dear Madam(s),
c. Dear Sirs(s)/Madam(s),

E.2.2 Variations on specified versions of component "Greeting"

1. Punctuation missing at the end of greeting expression

2. Nouns in greeting expression in low case

E.2.3 Specified correct well-formed versions of component "IntroYourself"

(82) a. My name is NAME

E.2.4 Variations on specified versions of component "IntroYourself"

1. Use of lower case in "my" allowed if used in the middle of a sentence

E.2.5 Specified correct well-formed versions of component "YourDept"

(83) a. I work in/for the Marketing Department

E.2.6 Variations on specified versions of component "YourDept"

1. Substitution of *Marketing Department* by another department name

2. Substitution of verb form *work* by other forms of the same verb

3. Substitution of prepositions *in* or *for* by different prepositions

4. Omission of determiner *the*

5. Omission of subject (*I*)

E.2.7 Specified correct well-formed versions of component "Course"

The two specified patterns are obtained from concatenating response parts in (84), (85) and (86).

(84) a. I would like to sign up to take/attend/do (...)

b. I would like to take/attend/do (...)

(85) a. (...) the course on/called/∅ TITLE from the Human Resources Department on DATE AND TIME

b. (...) the TITLE course from the Human Resources Department on DATE AND TIME

c. (...) the course on/called/∅ TITLE on DATE AND TIME

d. (...) the TITLE course on DATE AND TIME

e. (...) the course on/called/∅ TITLE

f. (...) the TITLE course

(86) a. If TITLE is *Business Communication* DAY AND TIME are *Monday, Wednesday and Thursday between 9am and 10am.*

b. If TITLE is *E-Commerce and E-Business* DAY AND TIME are *Monday and Wednesday between 1pm and 2:30pm.*

E.2.8 Variations on specified versions of component "Course"

1. Omission of determiners and/or prepositions
2. Substitutions of determiners and/or prepositions by other determiners/prepositions
3. Omission of one of the valid course titles
4. Blending of *Title* and *Day and Time* corresponding to two different coruses
5. Substitution of verb form would with other *modal verbs* or with *will*
6. Use of lower cases in one of the proper names in the response component

E.2.9 Specified correct well-formed versions of component "Schedule"

(87) a. It does not affect my schedule as I am free on *Day and Time.*

b. The time and day for which the course is scheduled is perfect for me.

E.2.10 Variations on specified versions of component "Schedule"

1. Wrong use of negation: *no* instead of *not* or *affects not/no affects* instead of *does not affect*
2. Omission of subject (it, I)
3. Wrong subject-predicate order
4. Omission of determiner or preposition
5. Substitution of determiner or preposition by other determiners/prepositions

E.2.11 Specified correct well-formed versions of component "AuthorisedBy"

(88) a. I already have/got the authorisation to participate from/of the/my boss/head/manager
b. I already have/got the authorisation from/of the/my boss/head/manager

(89) a. (...) with the permission/authorisation from/of my/the boss/head/manager

E.2.12 Variations on specified versions of component "AuthorisedBy"

1. Wrong subject-predicate order
2. Wrong position of adverb
3. Omission of determiner or preposition
4. Substitution of determiner or preposition by other determiners/prepositions

E.2.13 Specified correct well-formed versions of component "UsefulFuture"

(90) a. (...) to foster / in fostering my career/job/work
b. (...) to improve / in improving my knowledge/skills

E.2.14 Variations on specified versions of component "UsefulFuture"

1. Use of infinitive or gerundive with the wrong preposition
2. Omission of determiner or preposition
3. Substitution of determiner or preposition by other determiners/prepositions

E.2.15 Specified correct well-formed versions of component "FutureInterest"

The two specified patterns are obtained from concatenating response parts in (91) and(92).

(91) a. In the future I will/would/'d/might take/do/attend/visit/register for/sign up for (...)
b. In the future I might be taking/attending/participating in/registering for (...)

(92) a. (...) the TITLE course
b. (...) the course/courses on TITLE

E.2.16 Variations on specified versions of component "FutureInterest"

1. Use of infinitive or gerundive with the wrong preposition
2. Omission of determiner or preposition
3. Substitution of determiner or preposition by other determiners/prepositions

E.2.17 Specified correct well-formed versions of component "ComplClose"

(93) a. Best regards,
b. Yours faithfully/sincerely,

E.2.18 Variations on specified versions of component "ComplClose"

1. Punctuation missing at the end of complimentary closure
2. Substitution of *Yours* by *Your*

E.2.19 Specified correct well-formed versions of component "Signature"

(94) a. NAME
b. NAME SURNAME

E.2.20 Variations on specified versions of component "Signature"

1. Addition of unnecessary punctuation
2. Use of lower case in name or surname

E.3 NLP specifications for Task *Expresa tu satisfacción o insatisfacción con el producto* Smint

E.3.1 Specified correct well-formed versions of component "Saludo"

(95) a. Estimado/Apreciado Señor:
b. Muy señores míos:

E.3.2 Variations on specified versions of component "Saludo"

1. Punctuation missing at the end of complimentary closure
2. Use of *Querido* instead of *Estimado/Apreciado*

E.3.3 Specified correct well-formed versions of component "RazonCarta"

(96) a. Les escribo para darles mi opinión sobre *Smint*.

E.3.4 Variations on specified versions of component "RazonCarta"

1. Use informal forms (*le, darle*) instead of formal ones
2. Wrong verb form choices

E.3.5 Specified correct well-formed versions of component "Opinion"

(97) a. Creo que (...)
b. Yo creo que (...)
c. A mi modo ver (...)
d. Para mi (...)
e. me gusta NP
f. me gusta mucho NP
g. no me gusta NP
h. no me gusta nada NP

E.3.6 Variations on specified versions of component "Opinion"

1. Substitution of specified prepositions by other prepositions
2. Omission of pronoun *me* in sentences where needed

E.3.7 Specified correct well-formed versions of component “MasInfo”

(98) a. podría(n) decirme si
b. quería saber si
c. me gustaría saber si

E.3.8 Variations on specified versions of component “MasInfo”

1. Use of tenses other than present indicative in the completive sentence introduce by *si*

2. Check use of singular/plural form coherent with use of singular/plural form in *Saludo*

E.3.9 Specified correct well-formed versions of component “Despedida”

(99) a. Un saludo,
b. Saludos cordiales,
c. Atentamente,

E.3.10 Variations on specified versions of component “Despedida”

1. Punctuation missing at the end of complimentary closure

E.3.11 Specified correct well-formed versions of component “Firma”

(100) a. NAME
b. NAME SURNAME

E.3.12 Variations on specified versions of component “Firma”

1. Addition of unnecessary punctuation

2. Use of lower case in name or surname

Appendix F

TAF and RIF analysis of the E.T. activity

F.1 TAF analysis

Description The learner is expected to watch and report the events happening in the film fragment.

Focus Meaning.

Outcome Report (fragmented).

Processes Understanding films in English; reporting events and actions in films.

Input The activity provides a video fragment of the relevant scenes in the movie *E. T. The Extra Terrestrial.* Each item includes a prompt.

Response type Limited production response

Teaching goal Pre-communicative learning

Assessment Formative

F.2 RIF analysis

E.T. – THE EXTRATERRESTRIAL

PROMPT	[None]
INSTRUCTIONS	Watch the fragment of the movie *E. T. – The Extraterrestrial* and respond to the comprehension questions.
INPUT DATA	
	Video of the corresponding fragment http://www.youtube.com/watch?v=e6Jm6P26S2A&NR=1
INPUT DATA	1) How did E.T. learn to speak English?
RESPONSE	
	(...)

Table F.1: RIF characterisation of the activity "E.T. – The Extraterrestrial".

N.B. The URL to the video does not seem to work in all countries. For this reason, I have a copy of the video locally which is a screen capture of this video. If you are interested in watching it please email me.

Appendix G

Teacher traning material in ICE3

G.1 Table for the characterisation of CALL/ICALL activities

This is the table teachers used to characterise and classify CALL activities according to their correction technique. This decision made the activity to be authored using Hot Potatoes or the AutoTutor Toolkit for Activity Creation.

Activity	Skill activity and study focus	CALL technique						
		AutoTutor	Hot Potatoes					
		Short answer	Multiple choice	Short answer	Match	Fill-the-gap	Crosswords	Other
1								
2								
3								
(...)								

Table G.1: Table for teachers to characterise their learning activities.

Appendix H

Formal analysis of the ICALL activities authored by teachers

H.1 Teacher 1's work plan

The following three pages are a PDF version of the work plan produced by T1 during the experiment described in Chapter 11.

CHEMISTRY Physical and chemical changes. Chemical reactions				
Activities and writing (folder) / **Laboratory**	**Computers activity**	SCIENCE 3rd of ESO		
Sessions	**Meeting point**	**Time**	**Resources**	
Session 1: Class and Computers activity SCIENCE TOPIC: Physical and chemical changes Teacher explains at the blackboard the differences between physical and chemical changes. Students write down to include in their folders. Then the whole class watch the Video: "Physical and chemical changes" and put in common some examples to tell them apart. Then the class meet in computers room in order to do the Hot Potatoes activities related to the video. - Physical and chemical changes Match	In class In computers room	2 hours	Computers Internet connection	
Session 2. Laboratory SCIENCE TOPIC: Physical changes. Sublimation. The sublimation of Iodine.	Laboratory	1 hour	Laboratory instrumental Iodine	
Session 3, 4. In class and in computers room. SCIENCE TOPIC: Chemical reactions. Reactants and products. Teacher explains at the blackboard examples of chemical reactions, types and the differences between reactants and products. Students write down to include in their folders. Next students do the Hot Potatoes activities: - Chemical reactions Quiz - Chemical reactions Cloze - Chemical reactions Cross File	In class In computers room	2 hours	Computers Internet connection	

Session 5. Laboratory SCIENCE TOPIC: Chemical reactions. The catalysts. The catalyst. The influence of Manganese dioxide to decompose Peroxide water.	Laboratory	1 hour	Laboratory instrumental Peroxide water Manganese dioxide
Session 6. In computers room. SCIENCE TOPIC: Chemical reactions. Reactants and products. KEY LANGUAGE: Active and passive form Students do the Autotutor activity: - Chemical reactions examples	In computers room At home	1 hour	Computers Internet connection
Session 7, 8 In class and in computers room. SCIENCE TOPIC: Chemical reactions. Stequiometry. First the teacher explains and prepares exercises to solve in class about the balance of the chemical reactions and the yield. Then the exercises are corrected and put in common in class. Students do the Hot Potatoes activity: - Chemical reactions equilibrium Cloze - Chemical reactions yield Quiz	In class In computers room At home	2 hours	Computers Internet connection Paper, pencil case and calculator
Session 10 In computers room. SCIENCE TOPIC: Chemical reaction yields and rates. KEY LANGUAGE: 1st and 2nd condicional First the teacher review the 1st and the 2nd conditional. Students do the Autotutor activity: - Chemical reactions. Changing the rates.	In class In computers room	1 hour	Computers Internet connection

Session 11 In computers room. SCIENCE TOPIC: Analysing graphs. KEY LANGUAGE: Comparatives. In class let's review the way to analyse a graph by using comparatives and so on to relate the variables. Students do the Hot Potatoes activity: - Analysis graphs quiz. Students do the Autotutor activity: - Analysing graphs.	In class In computers room	2 hours	Computers Internet connection

H.2 Chemical reactions – Describing reactants and products

Figure H.1 shows a screen capture of the Activity *Chemical reactions – Describing reactants and products* including the instructions and its first item. This activity was created by Teacher 1. It is a CLIL activity for learners in the 3rd year of secondary education in a Catalan school.

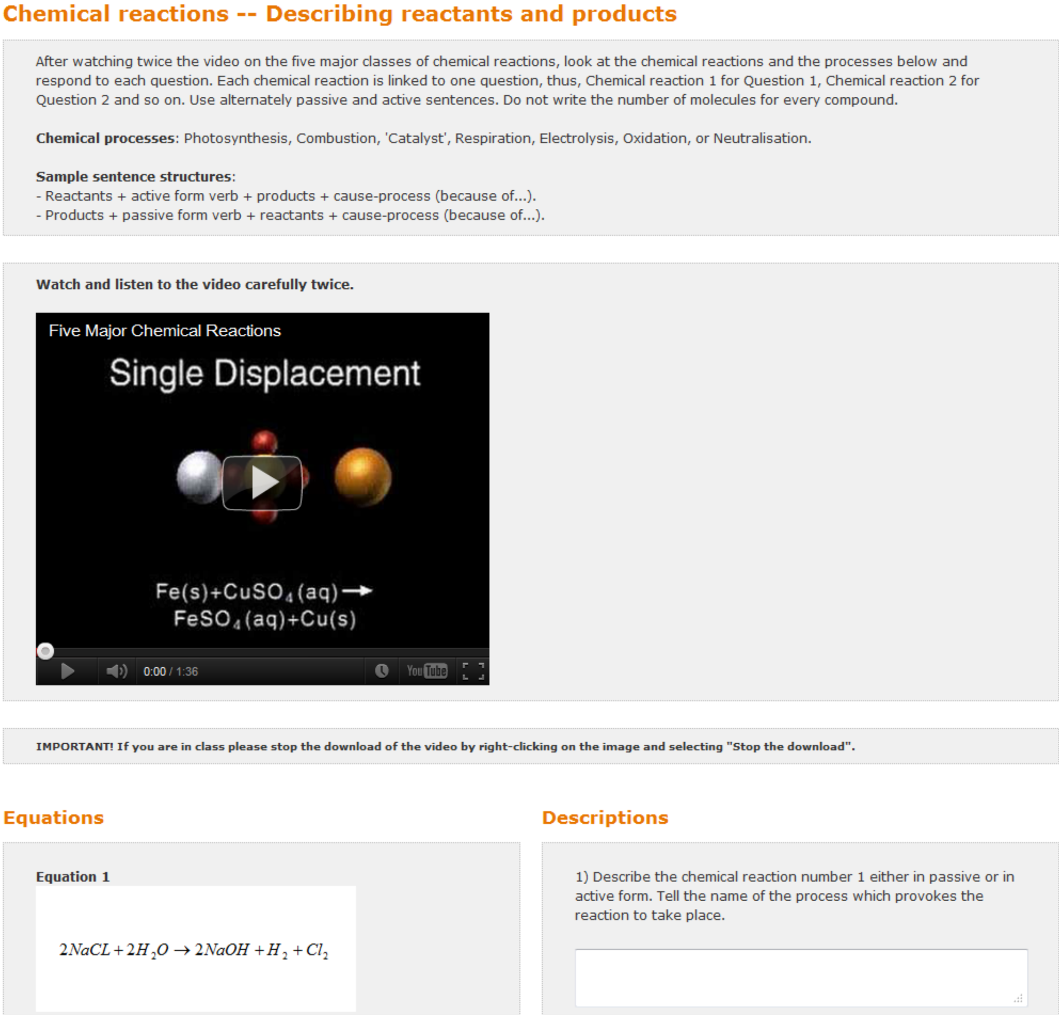


Figure H.1: Screen capture of the activity "Chemical reactions – Reactants and Products" created by T1.

H.2.1 TAF analysis

Description The learner is expected to read and describe the reactants, the products and the chemical process involved in a series of given chemical equations.

Focus Form/Reading chemical equations.

Outcome None.

Processes Understanding chemical equations in IUPAC nomenclature; use active and passive voice to describe the reactants, products, and process of chemical equations.

Input The activity provides a video explaining the different types of chemical reactions and a list of the seven types of chemical reactions. It also includes linguistic formulas to be used for the construction of sentences in active and passive voice. For each of the items, a chemical equation is provided.

Response type Limited production response

Teaching goal Two different types of goals:

- CLIL: Ability to read chemical equations
- FLTL: Pre-communicative learning

Assessment Formative

Chemical Reactions – Reactants and products

Prompt	[None]
Instructions	After watching twice the video on the five major classes of chemical reactions, look at the chemical reactions and the processes below and respond to each question. Each chemical reaction is linked to one question, thus, Chemical reaction 1 for Question 1, Chemical reaction 2 for Question 2 and so on. Use alternately passive and active sentences. Do not write the number of molecules for every compound.
Input data	(Thematic) **Chemical processes**: Photosynthesis, Combustion, 'Catalyst', Respiration, Electrolysis, Oxidation, or Neutralisation.
Input data	(Linguistic) Sample sentence structures: - Reactants + active form verb + products + cause-process (because of...). - Products + passive form verb + reactants + cause-process (because of...).
Input data	Watch and listen to the video carefully twice. http://youtu.be/tE4668aarck
Input data	1) Describe the chemical reaction number 1 either in passive or in active form. Tell the name of the process which provokes the reaction to take place. $2NaCL + 2H_2O \rightarrow 2NaOH + H_2 + Cl_2$
Response	
	(...)

Table H.1: RIF characterisation of the activity "Chemical reactions – Reactants and products" by Teacher 1.

H.2.2 RIF analysis

H.2.2.1 Detailed RIF analysis for Item 1

THEMATIC CONTENT OF THE EXPECTED RESPONSE	
ENTITIES	– Salt – Water – Hydrogen – Sodium hydroxide – Chlorine – Electrolysis
RELATIONS	– Reactants generating products – Electrolysis driving a reaction

LINGUISTIC CONTENT OF THE EXPECTED RESPONSE	
FUNCTIONAL	– Describe chemical reaction and *X and Y produce/generate/give P and Q due to Z* *P and Q are produced/generated/formed by P and Q because of Z*
SYNTACTIC	– Use passive or active voice
LEXICAL	– *water, salt, sodium chloride, electrolysis, give, produce, are formed by, due to, because of...*
PRAGMATICS	– Use capital letters in names of chemical elements.
GRAPHOLOGY	– Use the appropriate spelling.

Table H.2: Detailed RIF analysis for Item 1 in the Activity “Chemical reactions – Reactants and products”.

H.3 Chemical reactions – Calculating theoretical yields

Figure H.2 shows a screen capture of the Activity *Chemical reactions – Changing rates* including the instructions and its first two items. This activity was created by Teacher 1. It is a CLIL activity for learners in the 3rd year of secondary education in a Catalan school.

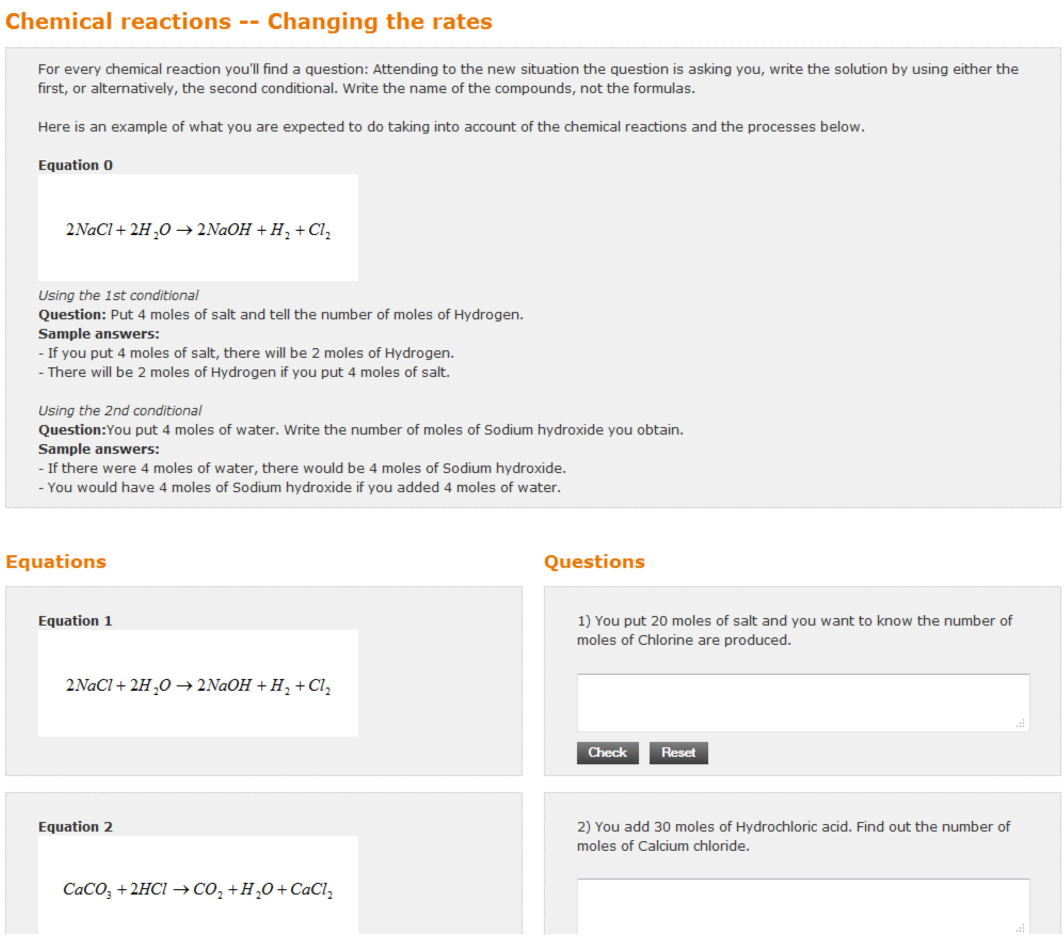


Figure H.2: Screen capture of the activity “Chemical reactions – Changing rates” created by T1.

H.3.1 TAF analysis

Description Learners are expected to interpret and word chemical equations including the ability to calculate the number of moles of products/reactants involved using conditional sentences.

Focus Form/Reading chemical equations

Outcome None

Processes Understanding chemical chemical equations in IUPAC nomenclature; describing in words the number of moles put and/or yielded in a chemical reaction; using first and second conditionals.

Input The activity provides an example of the what learners should do, including a chemical equation and some linguistic hints regarding the structure of the1st and 2nd conditional.

Response type Limited production response

Teaching goal Two different types of goals:

- CLIL: Ability to read chemical equations including yields and rates
- FLTL: Pre-communicative learning

Assessment Formative

H.3.2 RIF analysis

CHEMICAL REACTIONS – CHANGING RATES

PROMPT	[None]
INSTRUCTIONS	For every chemical reaction you'll find a question: Attending to the new situation the question is asking you, write the solution by using either the first, or alternatively, the second conditional. Write the name of the compounds, not the formulas.
EXAMPLE	Here is an example of what you are expected to do taking into account of the chemical reactions and the processes below. 0) Put 4 moles of salt and tell the number of moles of Hydrogen. $2NaCl + 2H_2O \rightarrow 2NaOH + H_2 + Cl_2$ Sample answers: **Using the 1st conditional** a) *If you put 4 moles of salt, there will be 2 moles of Hydrogen.* b) *There will be 2 moles of Hydrogen if you put 4 moles of salt.* **Using the 2nd conditional** a) *If there were 4 moles of water, there would be 4 moles of Sodium hydroxide.* b) *You would have 4 moles of Sodium hydroxide if you added 4 moles of water.*
INPUT DATA	1) You put 20 moles of salt and you want to know the number of moles of Chlorine are produced. $2NaCl + 2H_2O \rightarrow 2NaOH + H_2 + Cl_2$
RESPONSE	(...)

Table H.3: RIF characterisation of the activity "Chemical reactions – Changing rates" by Teacher 1.

H.3.2.1 Detailed RIF analysis for Item 1

Thematic content of the expected response	
Entities	– 20 moles of salt – 10 moles of Chlorine
Relations	– X moles of salt yield Y moles of Chlorine
Linguistic content of the expected response	
Functional	– Describe the theoretical yield of a reaction on the basis of its chemical equation using stoichiometry. – Expressing conditional events *if X* VERB$_{Pres}$*, then Y will* + INFINITIVE *if X* VERB$_{Past}$*, then Y would* + INFINITIVE
Syntactic	– Use passive or active voice
Lexical	– *salt, chlorine, moles, twenty, fifteen, have, add, yield, are formed, due to, because of, if, then...*
Pragmatics	– Use capital letters in names of chemical elements.
Graphology	– Use the appropriate spelling.

Table H.4: Detailed RIF analysis for Item 1 in the Activity "Chemical reactions – Changing rates".

H.4 Analysis of graphs (II)

Figure H.3 shows a screen capture of the activity *Analysing of graphs (II)* including the instructions and its first two items. This activity was created by Teacher 1. It is a CLIL activity for learners in the 3rd year of secondary education in a Catalan school.

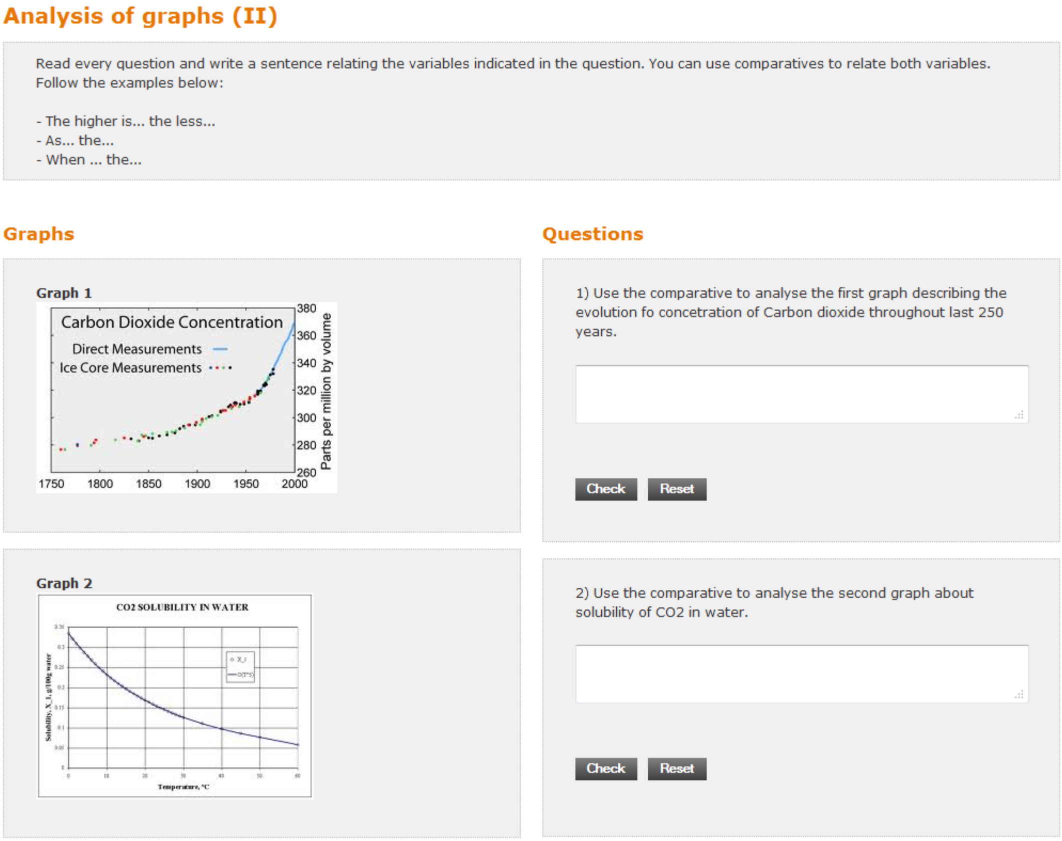


Figure H.3: Screen capture of the activity "Analysis of graphs (II)" created by T1.

H.4.1 TAF analysis

Description This activity focuses on the interpretation and verbalisation of phenomena reflected in graph-based representations containing two or three variables. Learners are expected to use linguistic structures for comparison.

Focus Form/Meaning

Outcome None

Processes Understanding graphs containing two and three variables; describing in words the evolution of variables in a graph; using comparison structures containing comparative adjectives and the corresponding adverbs or prepositions.

Input The activity includes some linguistic hints on how to use comparisons. The graphs are given as input to learners: They contain between two and three variables and a number of scales, labels and legends to be interpreted.

Response type Limited production response

Teaching goal Two different types of goals:

- CLIL: Ability to read graphs
- FLTL: Pre-communicative learning

Assessment Formative

Analysis of graphs (II)

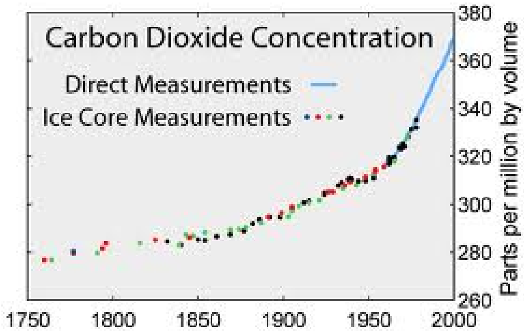
Prompt	[None]
Instructions	Read every question and write a sentence relating the variables indicated in the question. You can use comparatives to relate both variables.
Input data	(Linguistic) Follow the examples below: - The higher is... the less... - As... the... - When ... the...
Input data	1) Use the comparative to analyse the first graph describing the evolution of the concentration of Carbon dioxide throughout the last 250 years. 
Response	[ ] (...)

Table H.5: RIF characterisation of the activity "Chemical reactions – Changing rates" by Teacher 1.

H.4.2 RIF analysis

H.4.2.1 Detailed RIF analysis for Item 1

Thematic content of the expected response	
Entities	– Carbon, Carbon dioxide – time
Relations	– Time passes – The concentration of carbon dioxide grows over the years

Linguistic content of the expected response	
Functional	– Describe the evolution of a chemical or physical phenomenon over a period of time. – Expressing conditional events *as time passes, the* ADJECTIVE$_{Comp}$ *Y*
Syntactic	– Use comparative structures
Lexical	– *time, carbon, carbon dioxide, passes, flows, increases, concentration, mass...*
Pragmatics	– Use capital letters in names of chemical elements.
Graphology	– Use the appropriate spelling.

Table H.6: Detailed RIF analysis for Item 1 in the Activity “Analysis of graphs (II)”.

H.5 Daily routines II

Figure H.4 is a screen capture of the activity “Daily routines II” and its first three items. This activity was created by Teachers 2 and 3. It is an ESL activity for learners in the 1st and 2nd year of secondary education in a Catalan school.

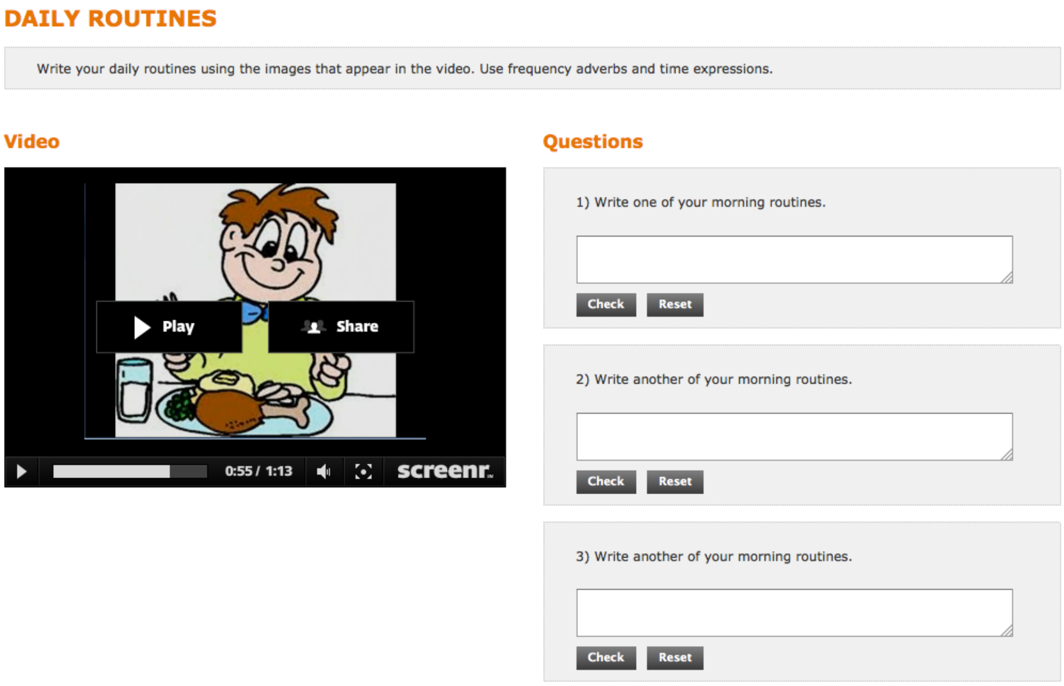


Figure H.4: Activity “Daily routines II” created by T2 and T3.

H.5.1 TAF analysis

Description Learners are required to write down their daily routines. The activity is thought as a preparation for a later oral presentation of one's daily routines.

Focus Form/Meaning

Outcome None

Processes Describing one's own routines; using the present simple tense; and using time expressions and/or frequency adverbs.

Input Learners are provided with a video showing images of people doing common actions and a prompt for each question. Instructions require them to use the present simple and time expressions.

Response type Limited production response

Teaching goal Pre-communicative learning

Assessment Formative

Daily routines II

Prompt	[None]
Instructions	Write your daily routines using the images that appear in the video. Use frequency adverbs and time expressions.
Input data	(Video) 
Input data	1) Write one of your morning routines.
Response	
Input data	2) Write another of your morning routines.
Response	
	(...)
Input data	4) Write one of your afternoon routines.
Response	
	(...)

Table H.7: RIF characterisation of the activity "Daily routines II" by Teachers 2 and 3.

H.5.2 RIF analysis

H.5.2.1 Detailed RIF analysis for Item 1

Thematic content of the expected response

Entities	– The person who does the action (first person singular pronoun, *I*) – Objects: book, comic, homework, bus... – Places: school, home, swimming pool... – People: brother, sister, father, grandmother...

Table H.8: Detailed RIF analysis for Item 1 in the Activity "Daily routines II" (continues).

Relations	– Passing from a sleep state to a state of consciousness – Leaving and returning to places – Intellectual, pedagogical and cognitive activities: read, play, watch... (...)
Linguistic content of the expected response	
Functional	– Describe actions performed during the day. – Describing the frequency with which actions take place – Using expressions of time
Syntactic	I VERB (OBJECT/PLACE/PERSON) adverb at MINUTES to/past HOUR – Use the present simple for first person singular – Word order in simple sentence including time expressions
Lexical	– *I, school, home, bus, teeth, shower, juice, milk, comic, TV, guitar, piano, play, swim, take, walk to...*
Pragmatics	– [None]
Graphology	– Use the appropriate spelling.

Table H.8: Detailed RIF analysis for Item 1 in the Activity “Daily routines II”.

H.6 The good and the bad student

Figure H.5 is a screen capture of the activity "The good and the bad student" and its first three items. This activity was created by Teachers 2 and 3. It is an ESL activity for learners in the 1st and 2nd year of secondary education in a Catalan school.

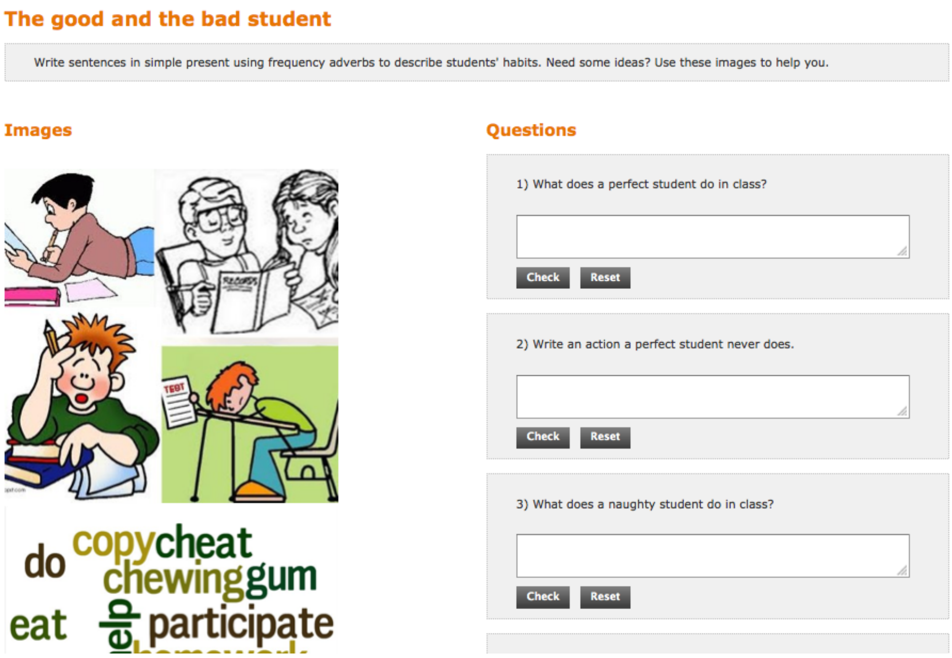


Figure H.5: Screen capture of the activity "The good and the bad student" created by T2 and T3.

H.6.1 TAF analysis

Description Learners are required to describe the habits of good and bad students, actions they typically do or do not do.

Focus Form

Outcome None

Processes Expressing habits from third parties; using the present simple tense; and using frequency adverbs.

Input Learners are provided with a word cloud including some verbs referring actions good and bad student might do, as well as some images for inspiration. Every item includes a question. Instructions require them to use the present simple and frequency adverbs.

Response type Limited production response

Teaching goal Pre-communicative learning

Assessment Formative

H.6.2 RIF analysis

The good and the bad student

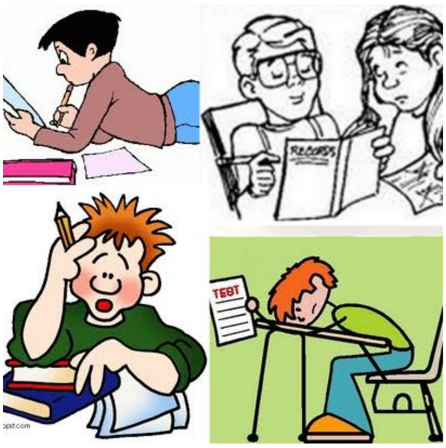
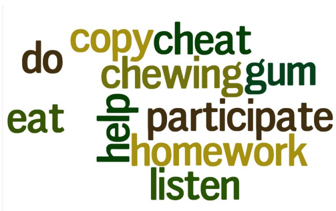
Prompt	[None]
Instructions	Write sentences in simple present using frequency adverbs to describe students' habits. Need some ideas? Use these images to help you.
Input data	 
Input data	1) What does a perfect student do in class?
Response	
Input data	2) Write an action a perfect student never does.
Response	
Input data	3) What does a naughty student do in class?
Response	
Input data	4) Write an action a naughty student never does.
Response	

Table H.9: RIF characterisation of the activity "The good and the bad student" by Teachers 2 and 3.

H.6.2.1 Detailed RIF analysis for Item 1

Thematic content of the expected response	
Entities	– the (good or bad) student, he, she – Objects: book, comic, homework... – Places: school, class... – People: teacher, classmates...
Relations	– Actions that are expected from a learner in class: pay attention, listen to the teacher, do the homework... – Things that are not expected from a learner in class: playing with other, copying...
Linguistic content of the expected response	
Functional	– Describe actions usually done or not done by third parties. – Describing the frequency with which actions take place
Syntactic	he/she ADVERB VERB (OBJECT/PLACE/PERSON) (in PLACE) – Use the present simple in third person singular: ending with -s – Word order in simple sentence including time expressions
Lexical	– *he, she, the (good/bad) student, school, home, homework, do, pay attention, listen to, play, talk...*
Pragmatics	– [None]
Graphology	– Use the appropriate spelling.

Table H.10: Detailed RIF analysis for Item 1 in the Activity "The good and the bad student".

H.7 Detailed ReSS specifications

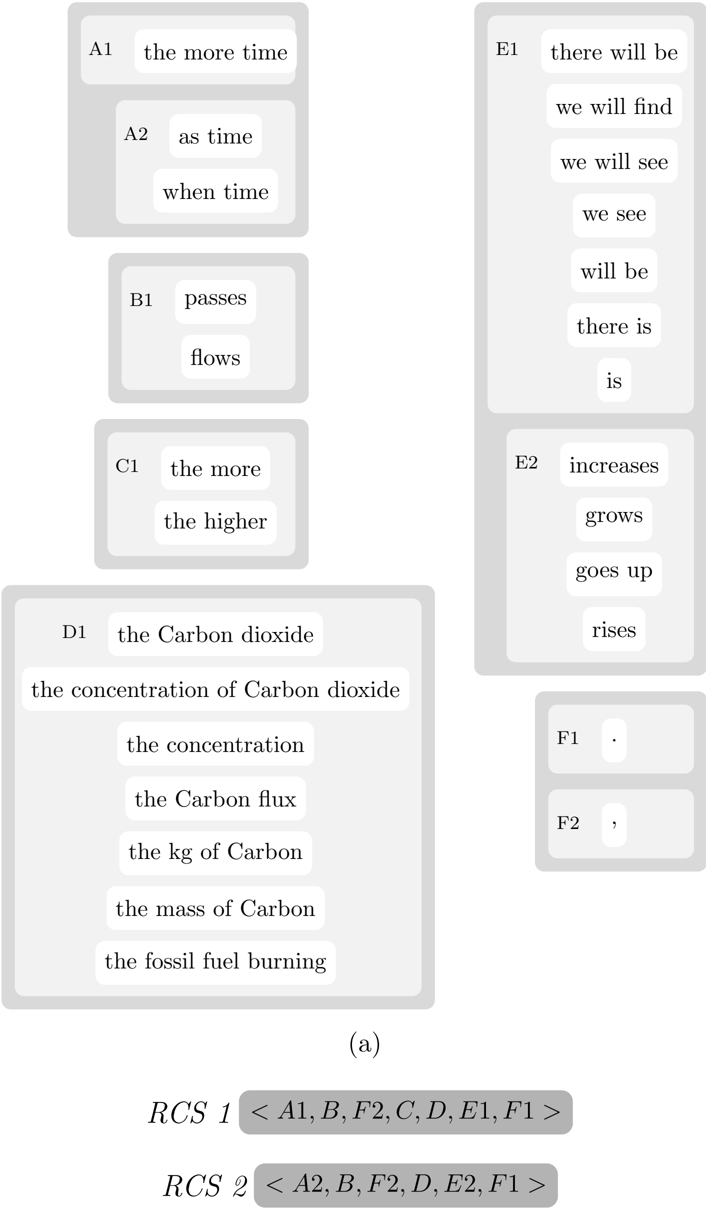
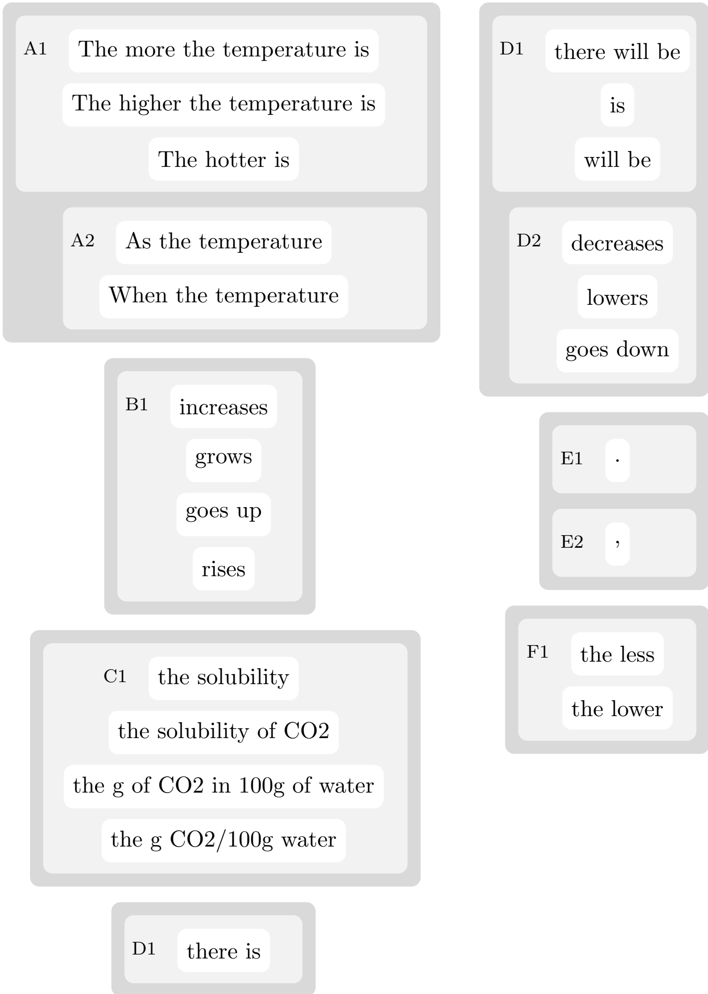


Figure H.6: ReSS specification RC and RCS for question 1 in activity AnalysisOfGraphs-v1 by T1.



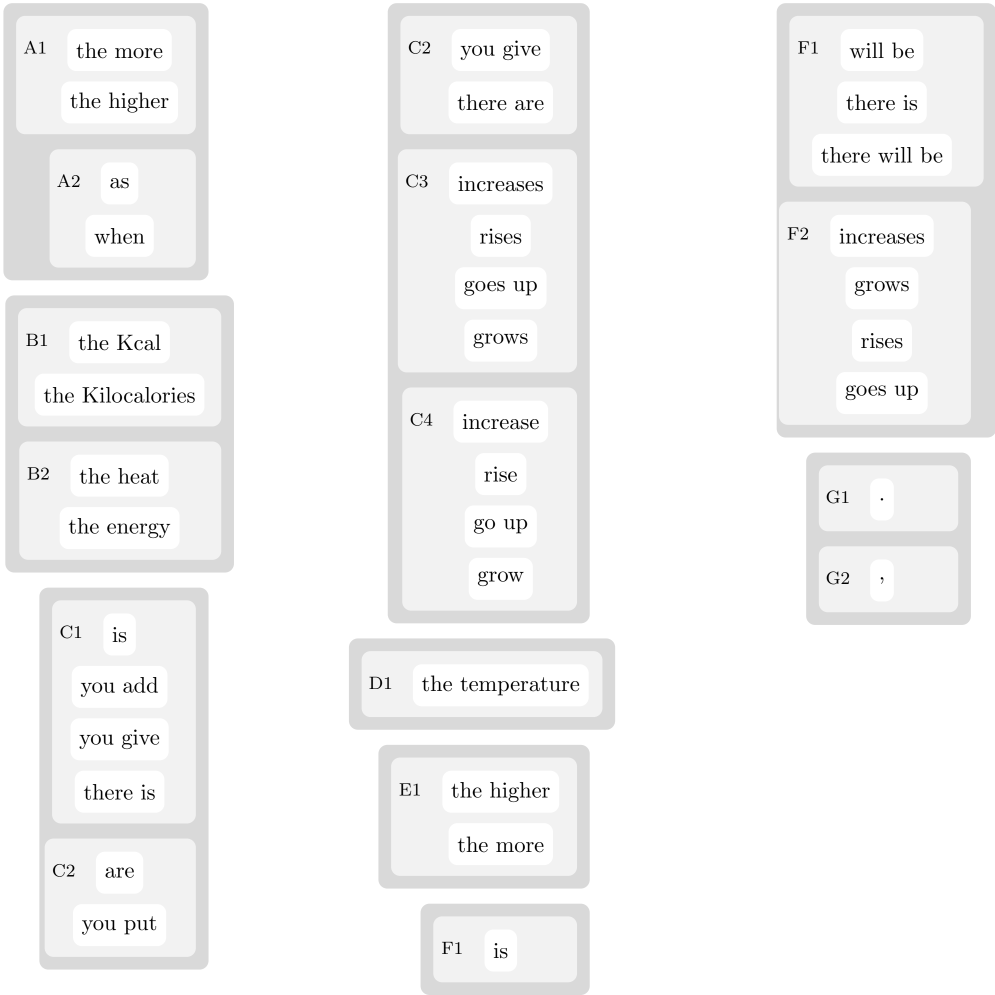
(a)

RCS 1 $< A1, E2, F, C, D1, E1 >$

RCS 2 $< A2, B, E2, C, D2, E1 >$

(b)

Figure H.7: ReSS specification RC and RCS for question 2 in activity AnalysisOfGraphs-v1 by T1.



(a)

RCS 1 $< A1, B1, C2, G2, E, D, F1, G1 >$

RCS 2 $< A1, B2, C1, G2, E, D, F1, G1 >$

RCS 3 $< A2, B1, C4, G2, D, F2, G1 >$

RCS 4 $< A2, B2, C3, G2, D, F2, G1 >$

(b)

Figure H.8: ReSS specification RC and RCS for question 3 in activity AnalysisOfGraphs-v1 by T1.

(a)

RCS 1 $< A1, B1, C1, F2, D1, E, F1 >$

RCS 2 $< A1, B2, C2, F2, D1, E, F1 >$

RCS 3 $< A2, B1, C3, F2, D1, E, F1 >$

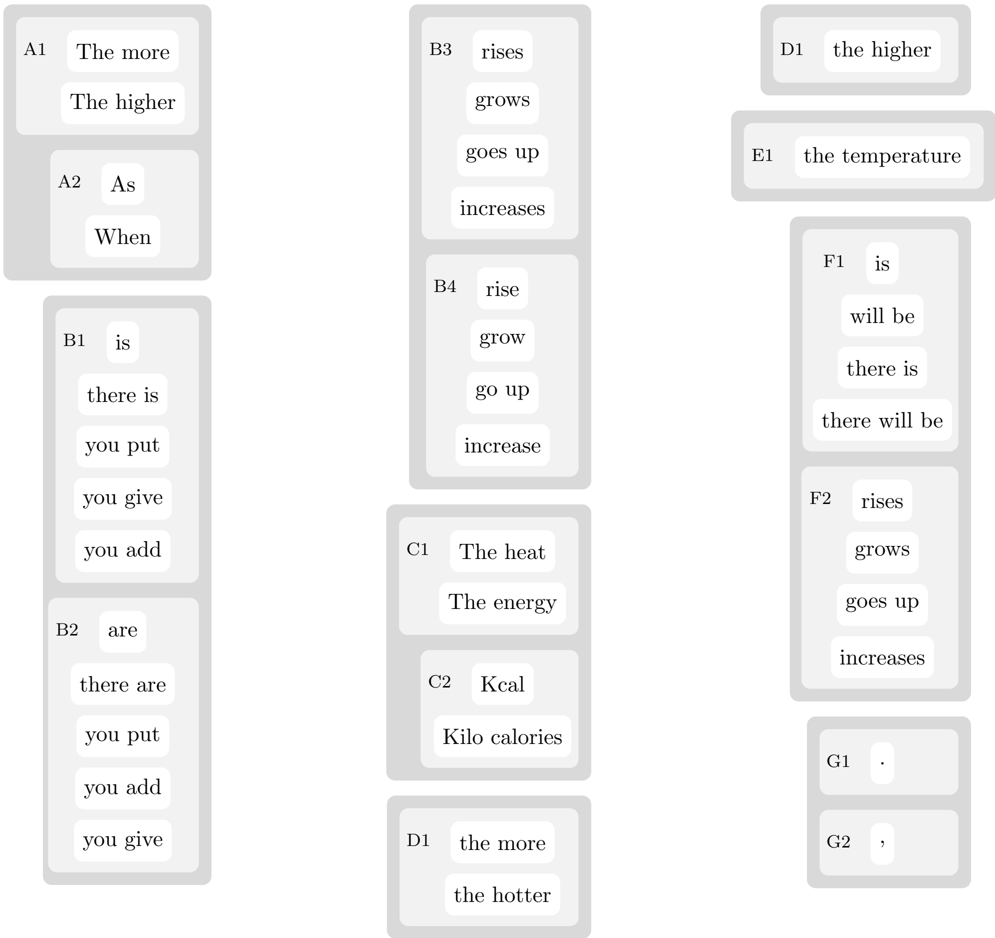
RCS 4 $< A2, B1, C3, F2, D2, F1 >$

RCS 5 $< A2, B2, C3, F2, D2, F1 >$

RCS 6 $< A2, B2, C4, F2, D1, E, F1 >$

(b)

Figure H.9: ReSS specification RC and RCS for question 4 in activity AnalysisOfGraphs-v1 by T1.



(a)

RCS 1 $< A1, C1, B1, G2, D, E, F1, G1 >$

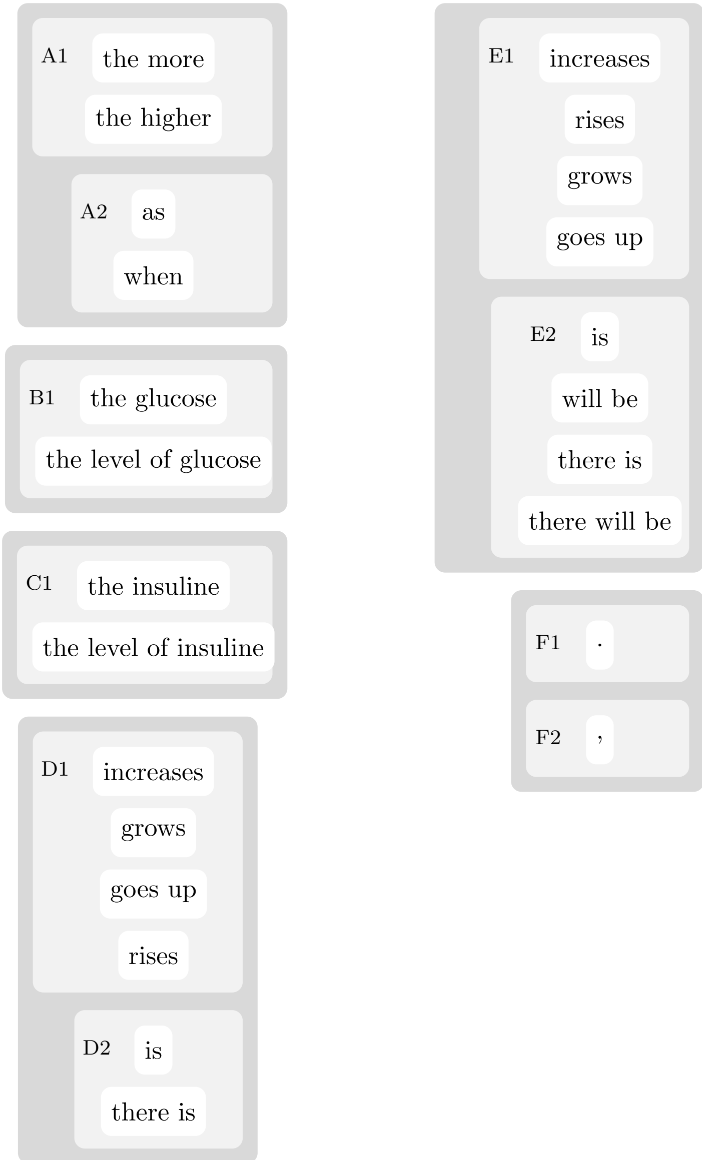
RCS 2 $< A1, C2, B2, G2, D, E, F1, G1 >$

RCS 3 $< A2, C1, B3, G2, E, F2, G1 >$

RCS 4 $< A2, C2, B4, G2, E, F2, G1 >$

(b)

Figure H.10: ReSS specification RC and RCS for question 5 in activity AnalysisOfGraphs-v1 by T1.



(a)

RCS 1 $< A1, B, D2, F2, A1, C, E2, F1 >$

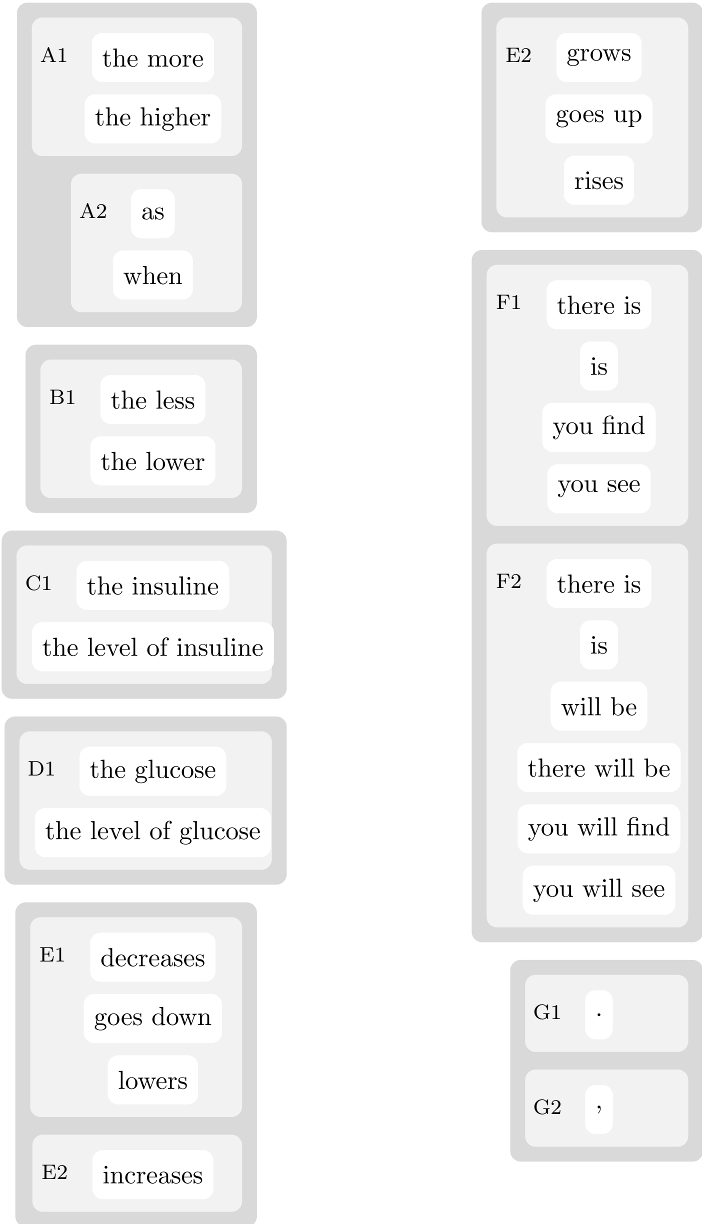
RCS 2 $< A1, C, D2, F2, A1, B, E2, F1 >$

RCS 3 $< A2, B, D1, F2, C, E1, F1 >$

RCS 4 $< A2, C, D1, F2, B, E1, F1 >$

(b)

Figure H.11: ReSS specification RC and RCS for question 6 in activity AnalysisOfGraphs-v1 by T1.



(a)

RCS 1 $< A1, C, F1, G2, B, D, F2, G1 >$

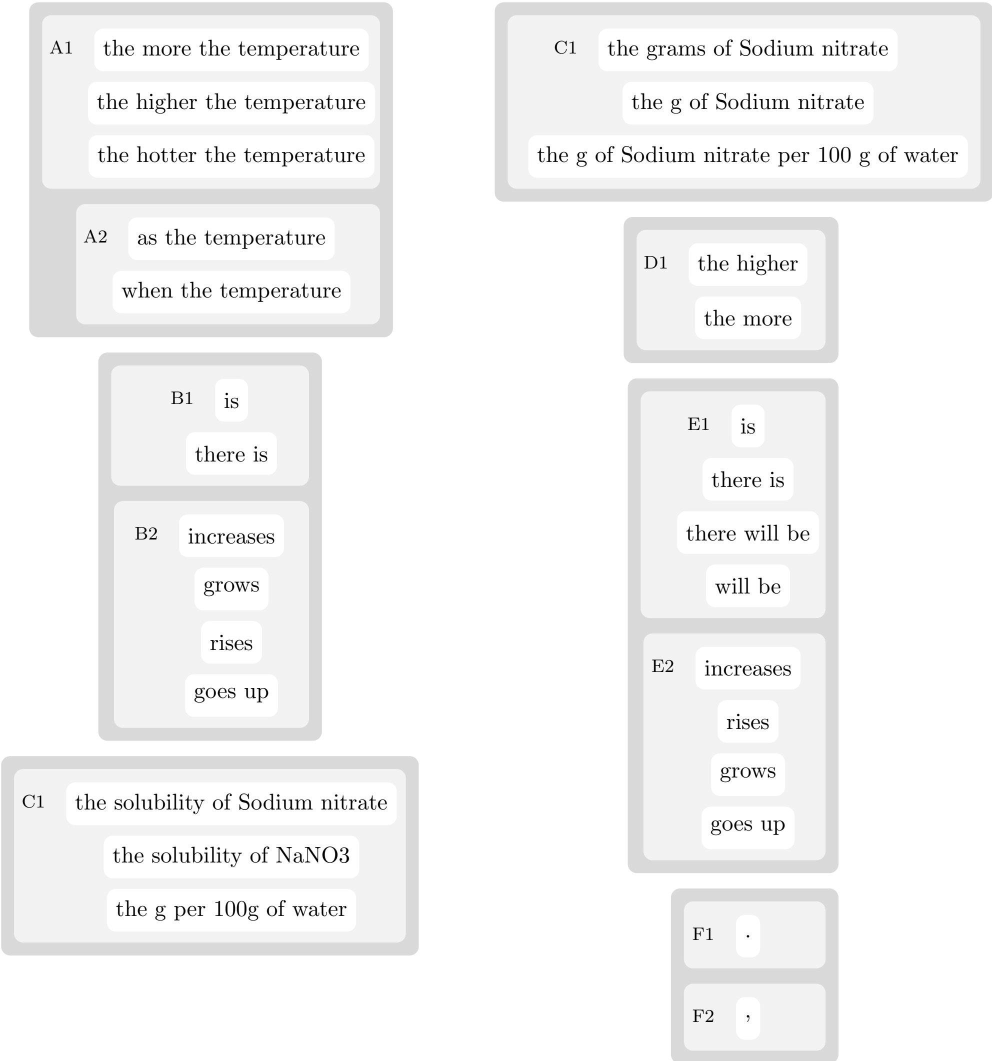
RCS 2 $< A2, C, E2, G2, D, E1, G1 >$

RCS 3 $< A2, D, E1, G2, C, E2, G1 >$

RCS 4 $< B, D, F1, G2, A1, C, F2, G1 >$

(b)

Figure H.12: ReSS specification RC and RCS for question 7 in activity AnalysisOfGraphs-v1 by T1.



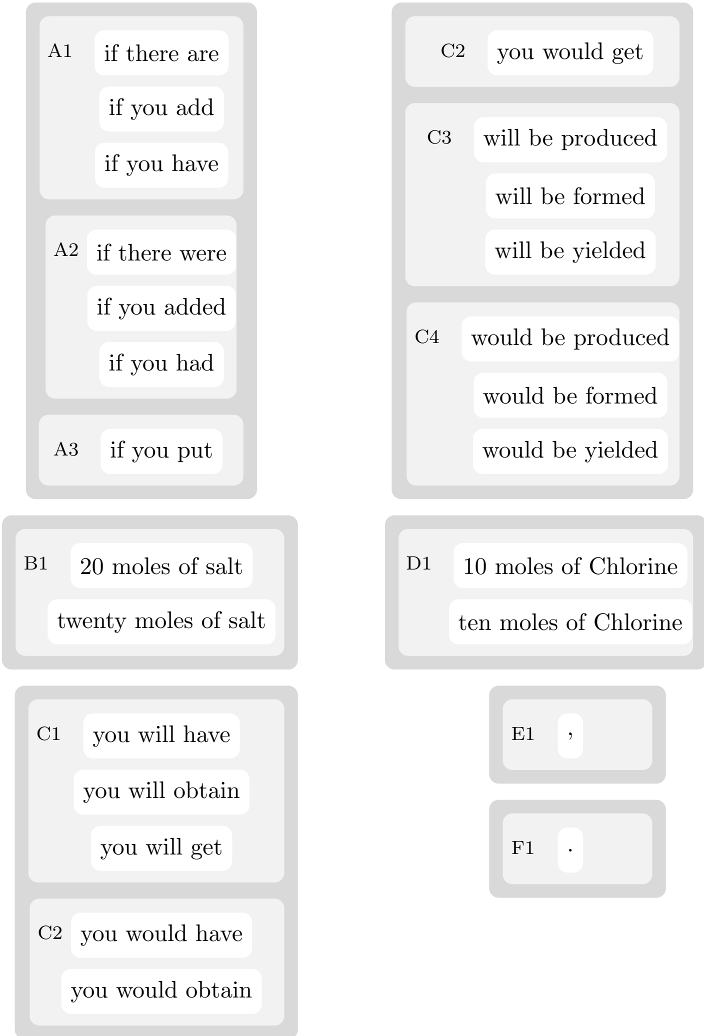
(a)

RCS 1 $< A1, B1, F2, D, C, E1, F1 >$

RCS 2 $< A2, B2, F2, C, E2, F1 >$

(b)

Figure H.13: ReSS specification RC and RCS for question 8 in activity AnalysisOfGraphs-v1 by T1.



(a)

RCS 1 $< A1, B, E, C1, D, F >$

RCS 2 $< A1, B, E, C1, D, F >$

RCS 3 $< A2, B, E, C2, D, F >$

RCS 4 $< A3, B, E, C1, D, F >$

RCS 5 $< A3, B, E, C2, D, F >$

RCS 6 $< C1, D, A1, B, F >$

RCS 7 $< C1, D, A3, B, F >$

RCS 8 $< C2, D, A2, B, F >$

RCS 9 $< C2, D, A2, B, F >$

RCS 10 $< C2, D, A3, B, F >$

RCS 11 $< D, C3, A1, B, F >$

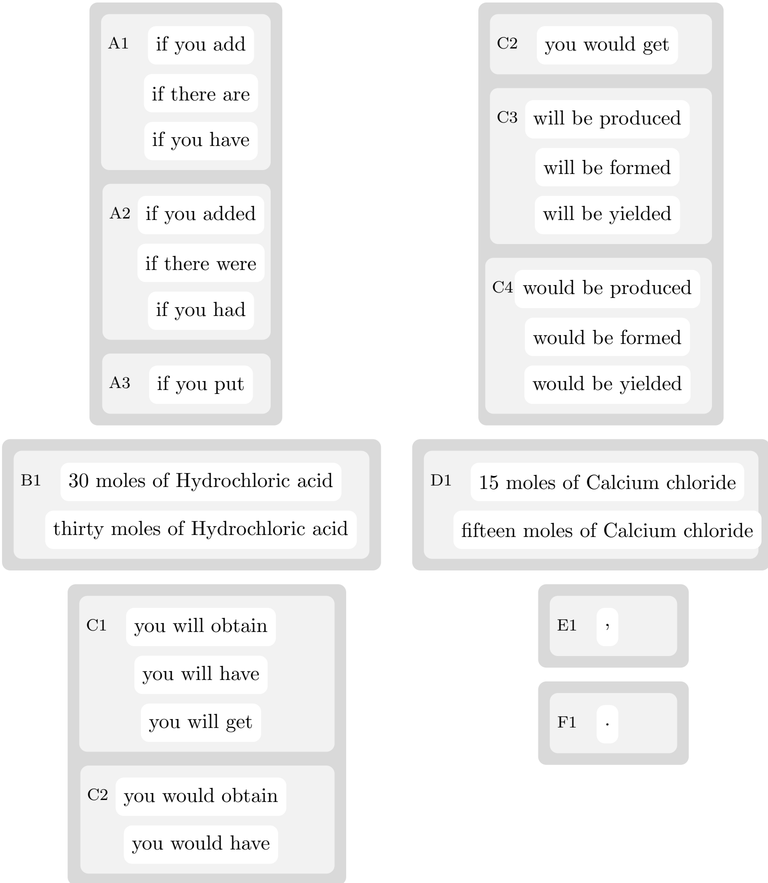
RCS 12 $< D, C3, A3, B, F >$

RCS 13 $< D, C4, A2, B, F >$

RCS 14 $< D, C4, A3, B, F >$

(b)

Figure H.14: ReSS specification RC and RCS for question 1 in activity Changingtherates1 by T1.



(a)

RCS 1 $< A1, B, E, C1, D, F >$

RCS 2 $< A2, B, E, C2, D, F >$

RCS 3 $< A3, B, E, C1, D, F >$

RCS 4 $< A3, B, E, C2, D, F >$

RCS 5 $< C1, D, A1, B, F >$

RCS 6 $< C1, D, A3, B, F >$

RCS 7 $< C2, D, A2, B, F >$

RCS 8 $< C2, D, A3, B, F >$

RCS 9 $< D, C3, A1, B, F >$

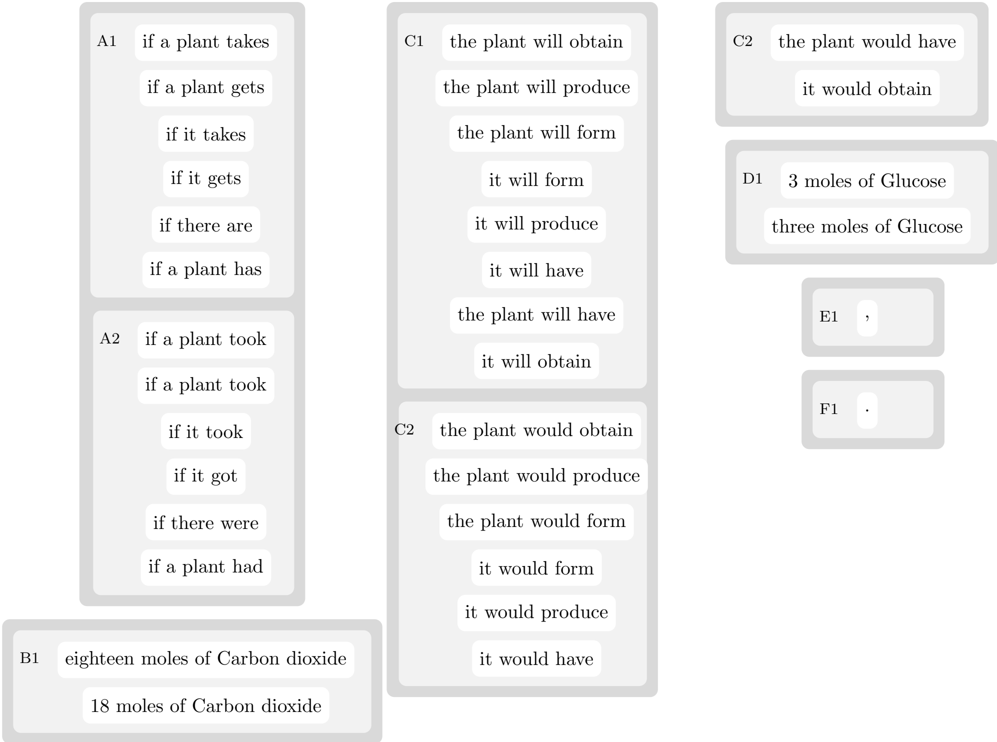
RCS 10 $< D, C3, A3, B, F >$

RCS 11 $< D, C4, A2, B, F >$

RCS 12 $< D, C4, A3, B, F >$

(b)

Figure H.15: ReSS specification RC and RCS for question 2 in activity Changingtherates1 by T1.



(a)

$RCS\ 1$ $< A1, B, E, C1, F, D >$

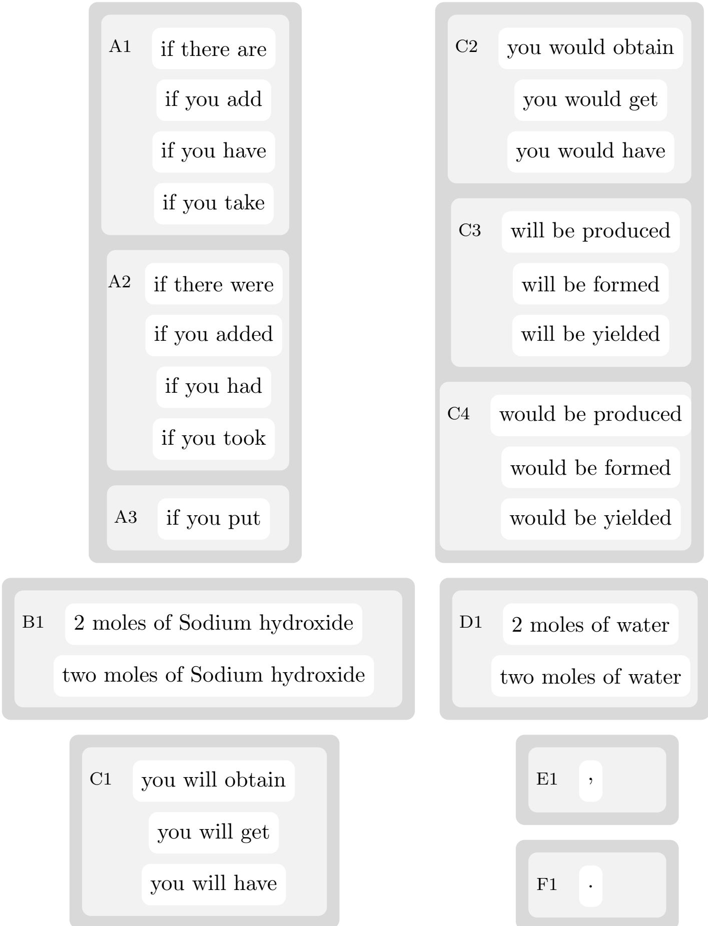
$RCS\ 2$ $< A2, B, E, C1, F >$

$RCS\ 3$ $< C1, D, A1, B, E, F >$

$RCS\ 4$ $< C2, D, A2, B, E, F >$

(b)

Figure H.16: ReSS specification RC and RCS for question 3 in activity Changingtherates1 by T1.



(a)

$RCS\ 1$ $< A1, B, E, C1, D, F >$

$RCS\ 2$ $< A2, B, E, C2, D, F >$

$RCS\ 3$ $< A3, B, E, C1, D, F >$

$RCS\ 4$ $< A3, B, E, C2, D, F >$

$RCS\ 5$ $< C1, D, A1, B, F >$

$RCS\ 6$ $< C1, D, A3, B, F >$

$RCS\ 7$ $< C2, B, A3, D, F >$

$RCS\ 8$ $< C2, D, A2, B, F >$

$RCS\ 9$ $< D, C3, A1, B, F >$

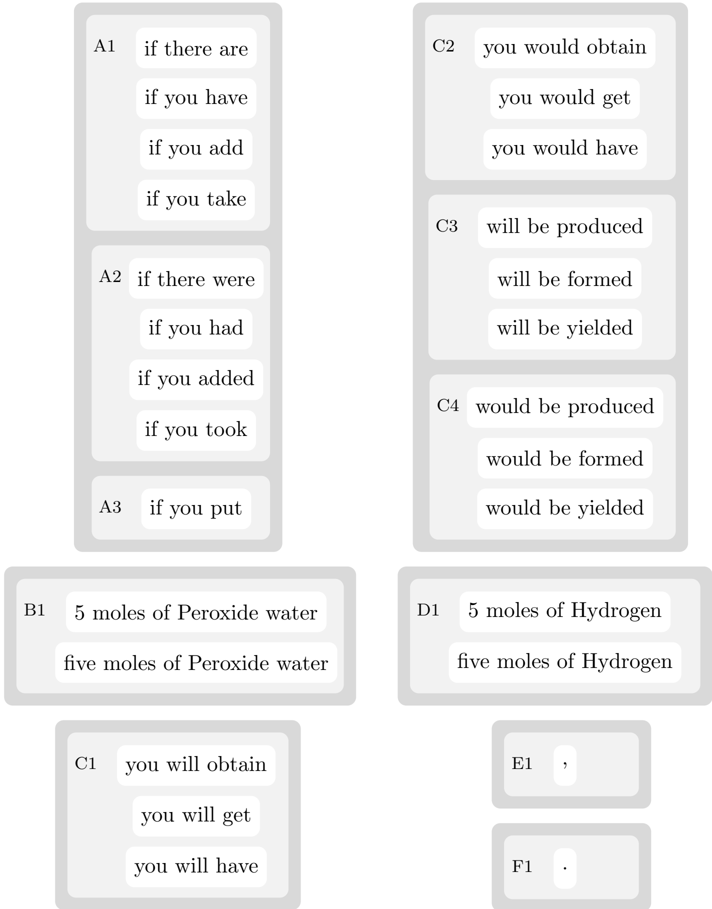
$RCS\ 10$ $< D, C3, A3, B, F >$

$RCS\ 11$ $< D, C4, A2, B, F >$

$RCS\ 12$ $< D, C4, A3, B, F >$

(b)

Figure H.17: ReSS specification RC and RCS for question 4 in activity Changingtherates1 by T1.



(a)

RCS 1 $< A1, B, E, C1, D, F >$

RCS 2 $< A2, B, E, C2, D, F >$

RCS 3 $< A3, B, E, C1, D, F >$

RCS 4 $< A3, B, E, C2, D, F >$

RCS 5 $< C1, D, A1, B, F >$

RCS 6 $< C1, D, A3, B, F >$

RCS 7 $< C2, D, A2, B, F >$

RCS 8 $< C2, D, A3, B, F >$

RCS 9 $< D, C3, A1, B, F >$

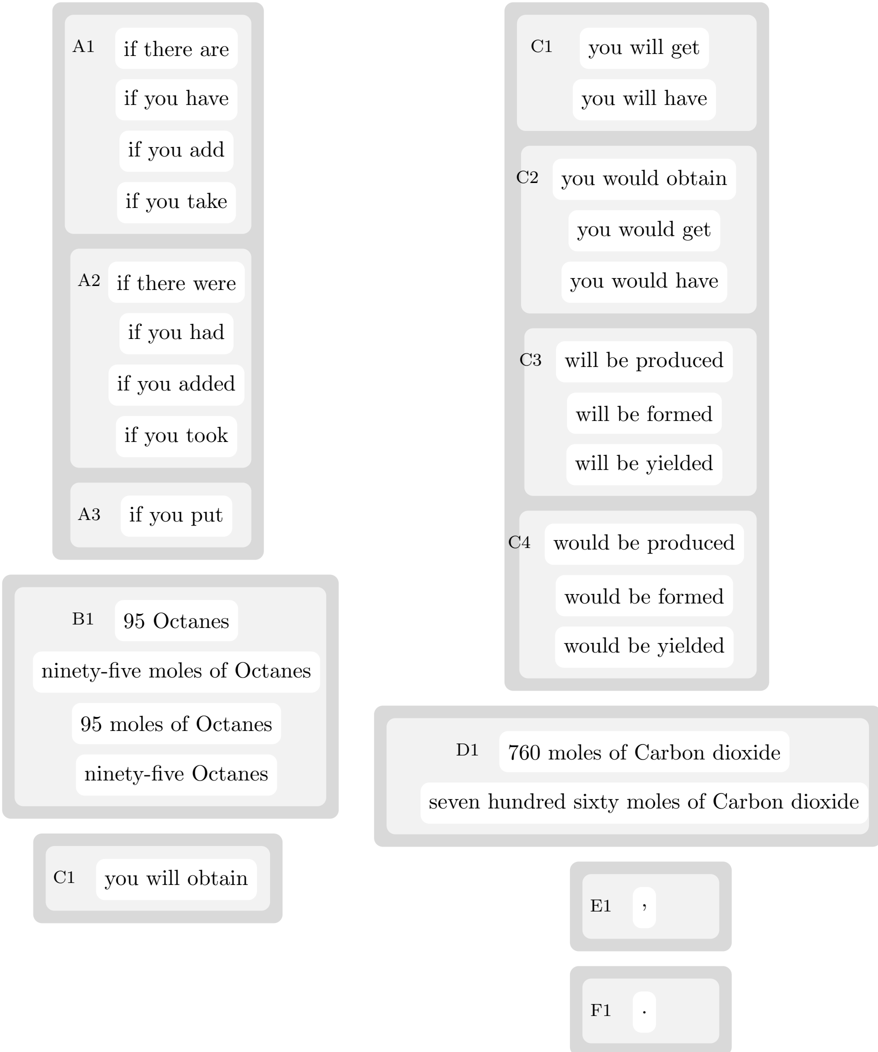
RCS 10 $< D, C3, A3, B, F >$

RCS 11 $< D, C4, A2, B, F >$

RCS 12 $< D, C4, A3, B, F >$

(b)

Figure H.18: ReSS specification RC and RCS for question 5 in activity Changingtherates1 by T1.



(a)

RCS 1 $< A1, B, E, C1, D, F >$ *RCS 7* $< C2, D, A2, B, F >$

RCS 2 $< A2, B, E, C2, D, F >$ *RCS 8* $< C2, D, A3, B, F >$

RCS 3 $< A3, B, E, C1, D, F >$ *RCS 9* $< D, C3, A1, B, F >$

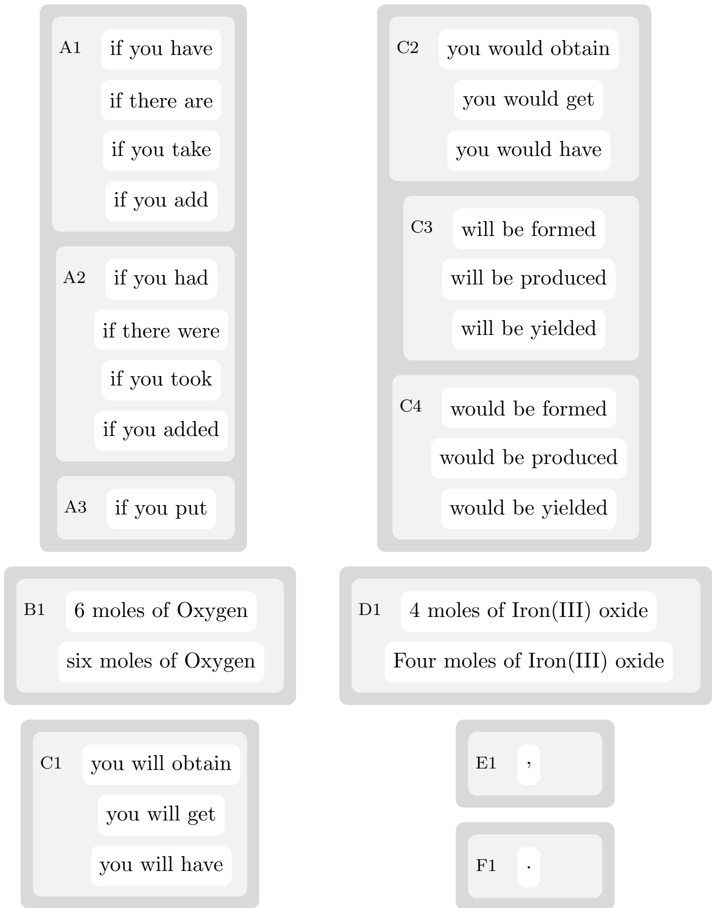
RCS 4 $< A3, B, E, C2, D, F >$ *RCS 10* $< D, C3, A3, B, F >$

RCS 5 $< C1, D, A1, B, F >$ *RCS 11* $< D, C4, A2, B, F >$

RCS 6 $< C1, D, A3, B, F >$ *RCS 12* $< D, C4, A3, B, F >$

(b)

Figure H.19: ReSS specification RC and RCS for question 6 in activity Changingtherates1 by T1.



(a)

RCS 1 $< A1, B, E, C1, D, F >$

RCS 2 $< A2, B, E, C2, D, F >$

RCS 3 $< A2, D, A3 >$

RCS 4 $< C1, D, A1, B, F >$

RCS 5 $< C2, D, A2, B, F >$

RCS 6 $< D, C3, A1, B, F >$

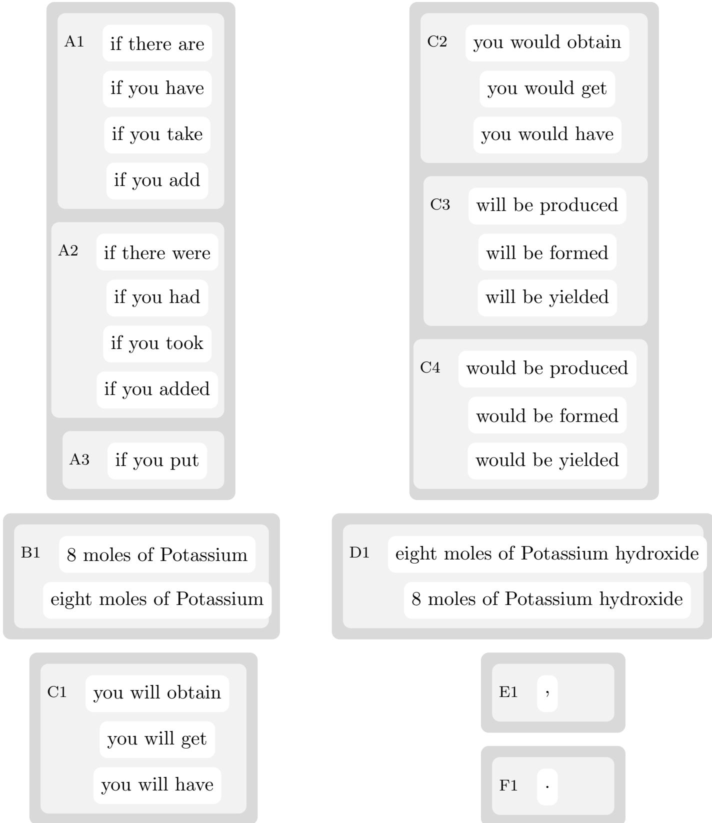
RCS 7 $< D, C3, A3, B, F >$

RCS 8 $< D, C4, A2, B, F >$

RCS 9 $< D, C4, A3, B, F >$

(b)

Figure H.20: ReSS specification RC and RCS for question 7 in activity Changingtherates1 by T1.



(a)

RCS 1 $< A1, B, E, C1, D, F >$

RCS 2 $< A2, B, E, C2, D, F >$

RCS 3 $< A3, B, E, C1, D, F >$

RCS 4 $< A3, B, E, C2, D, F >$

RCS 5 $< C1, D, A1, B, F >$

RCS 6 $< C1, D, A3, B, F >$

RCS 7 $< C2, D, A2, B, F >$

RCS 8 $< C2, D, A3, B, F >$

RCS 9 $< D, C3, A1, B, F >$

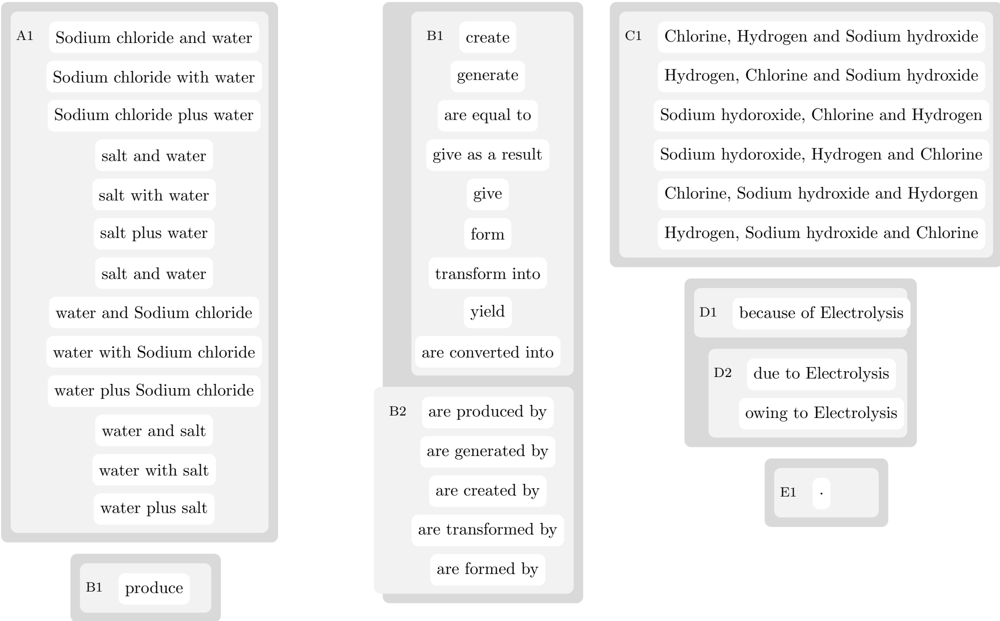
RCS 10 $< D, C3, A3, B, F >$

RCS 11 $< D, C4, A2, B, F >$

RCS 12 $< D, C4, A3, B, F >$

(b)

Figure H.21: ReSS specification RC and RCS for question 8 in activity Changingtherates1 by T1.



(a)

RCS 1 $< A, A, B1, C, C, C, D1 >$ *RCS 5* $< C, C, C, B2, A, A, D1 >$

RCS 2 $< A, A, B1, C, C, C, D1, E >$ *RCS 6* $< C, C, C, B2, A, A, D1, E >$

RCS 3 $< A, A, B1, C, C, C, D2 >$ *RCS 7* $< C, C, C, B2, A, A, D2 >$

RCS 4 $< A, A, B1, C, C, C, D2, E >$ *RCS 8* $< C, C, C, B2, A, A, D2, E >$

(b)

Figure H.22: ReSS specification RC and RCS for question 1 in activity ChemicalReactions-v1 by T1.

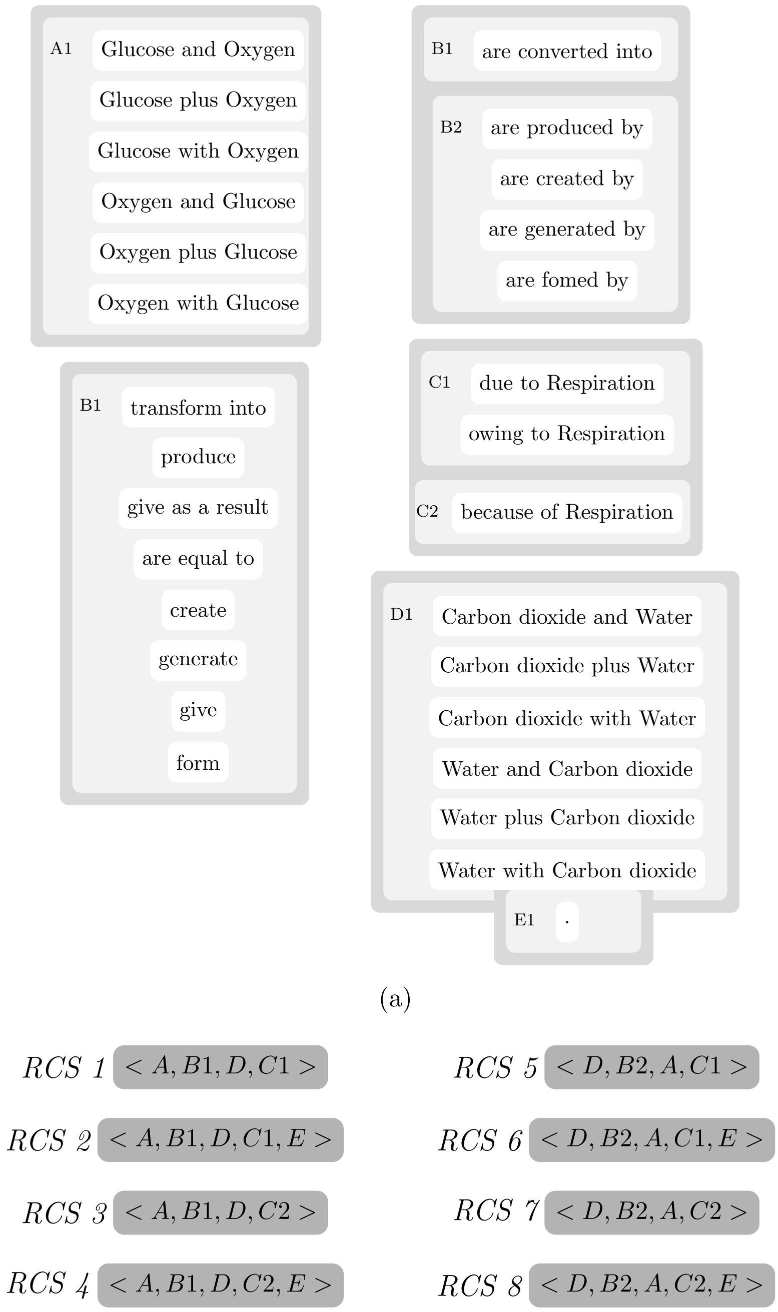
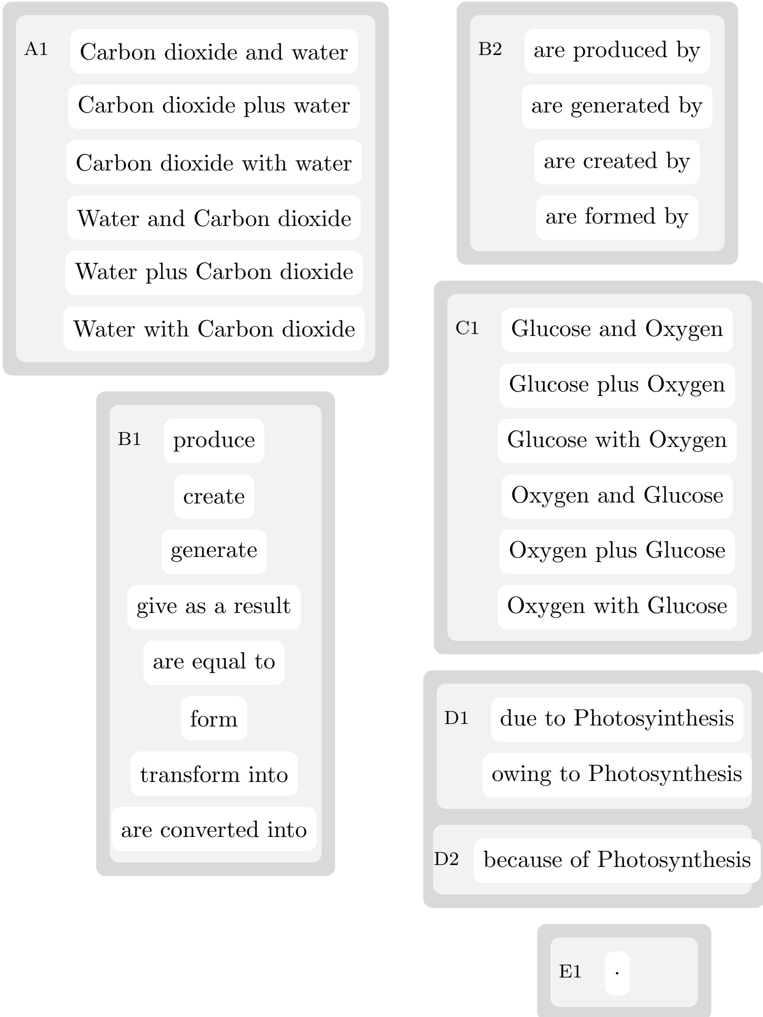


Figure H.23: ReSS specification RC and RCS for question 2 in activity ChemicalReactions-v1 by T1.



(a)

RCS 1 $< A, B1, C, D1, E >$

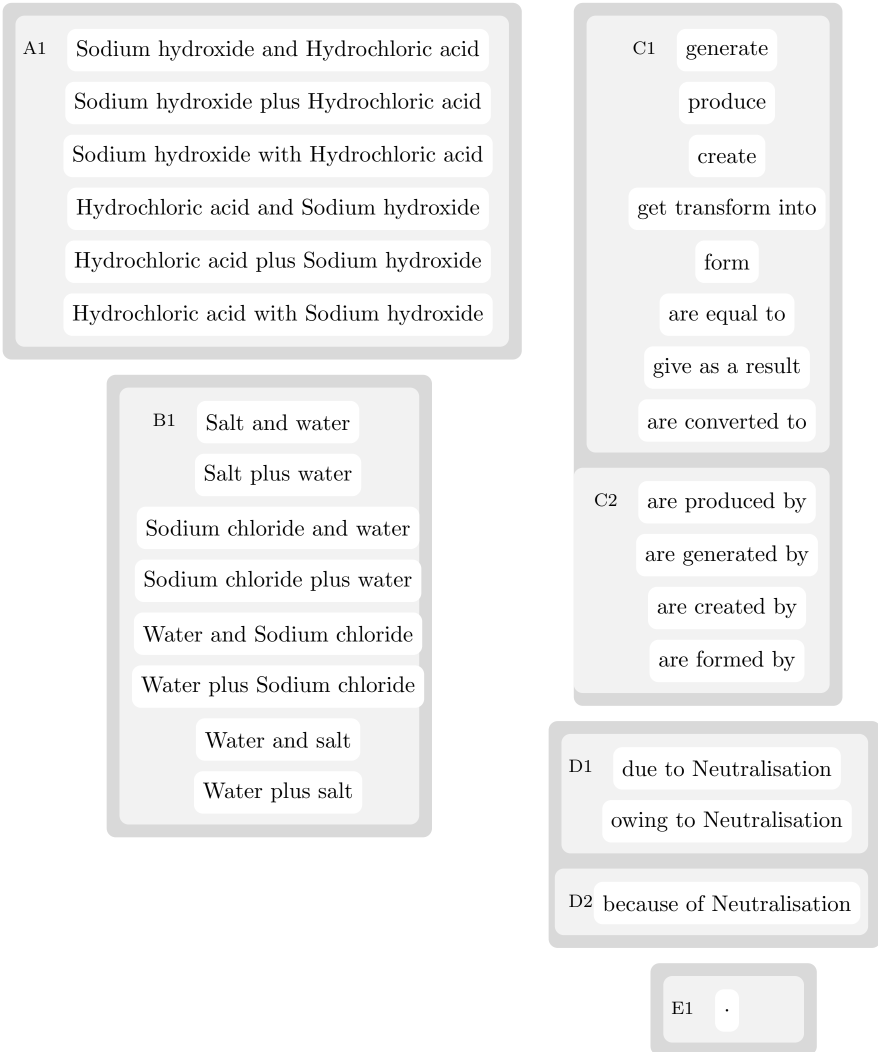
RCS 2 $< A, B1, C, D2, E >$

RCS 3 $< C, B2, A, D1, E >$

RCS 4 $< C, B2, A, D2, E >$

(b)

Figure H.24: ReSS specification RC and RCS for question 3 in activity ChemicalReactions-v1 by T1.



(a)

RCS 1 $< A, C1, B, D1, E >$

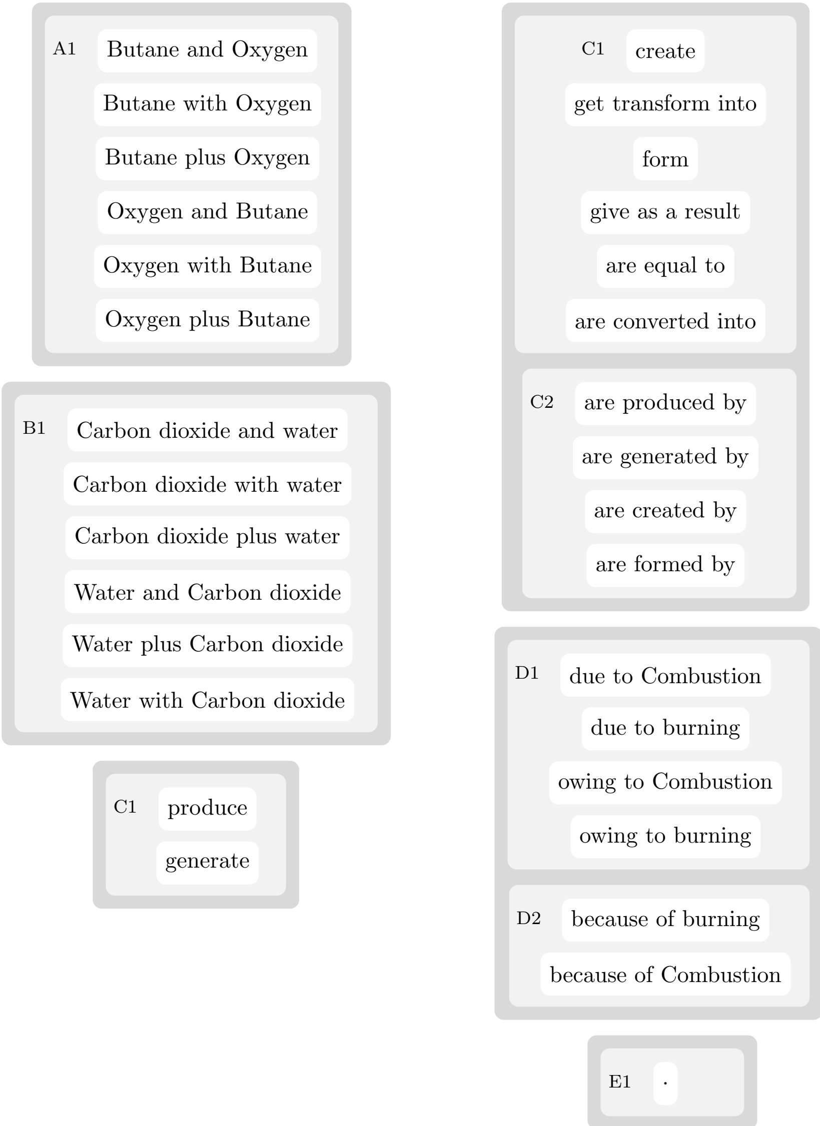
RCS 2 $< A, C1, B, D2, E >$

RCS 3 $< B, C2, A, D1, E >$

RCS 4 $< B, C2, A, D2, E >$

(b)

Figure H.25: ReSS specification RC and RCS for question 4 in activity ChemicalReactions-v1 by T1.



(a)

RCS 1 $< A, C1, B, D1, E >$

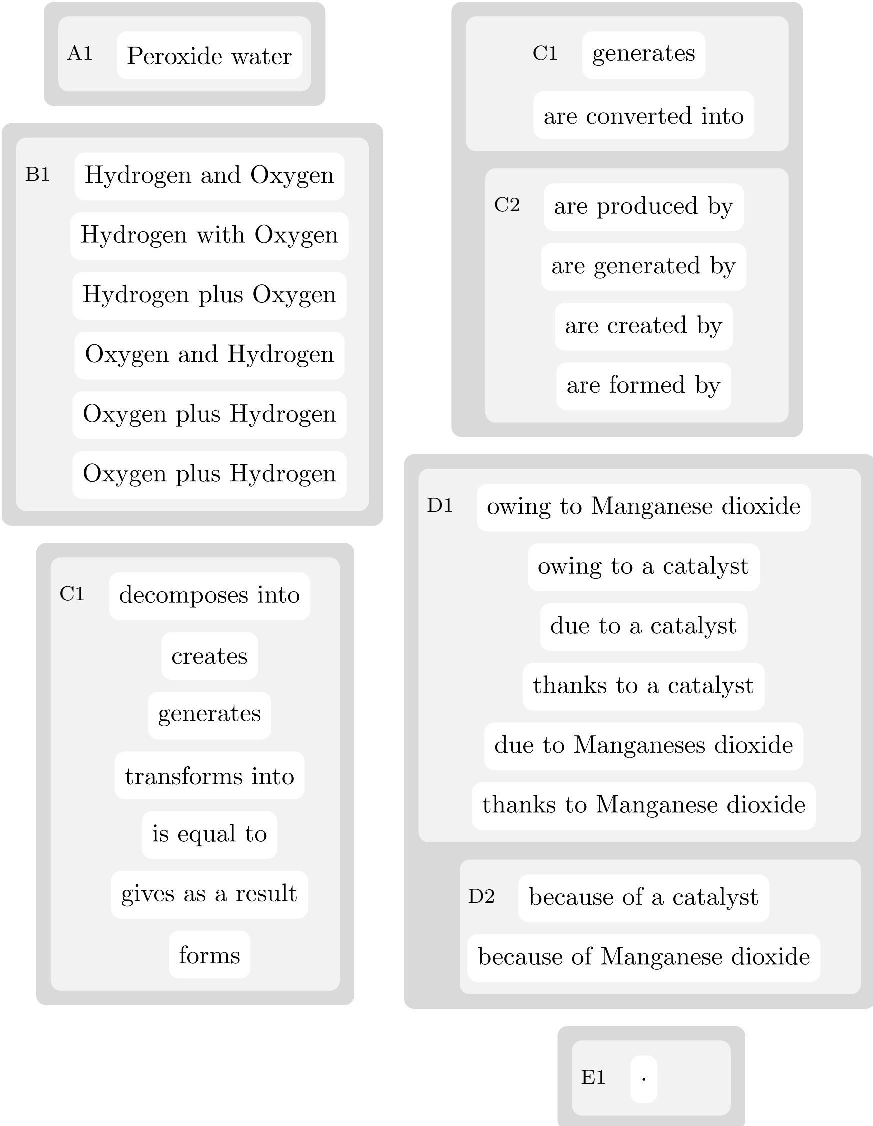
RCS 2 $< A, C1, B, D2, E >$

RCS 3 $< B, C2, A, D1, E >$

RCS 4 $< B, C2, A, D2, E >$

(b)

Figure H.26: ReSS specification RC and RCS for question 5 in activity ChemicalReactions-v1 by T1.



(a)

RCS 1 $< A, C1, B, D1, E >$

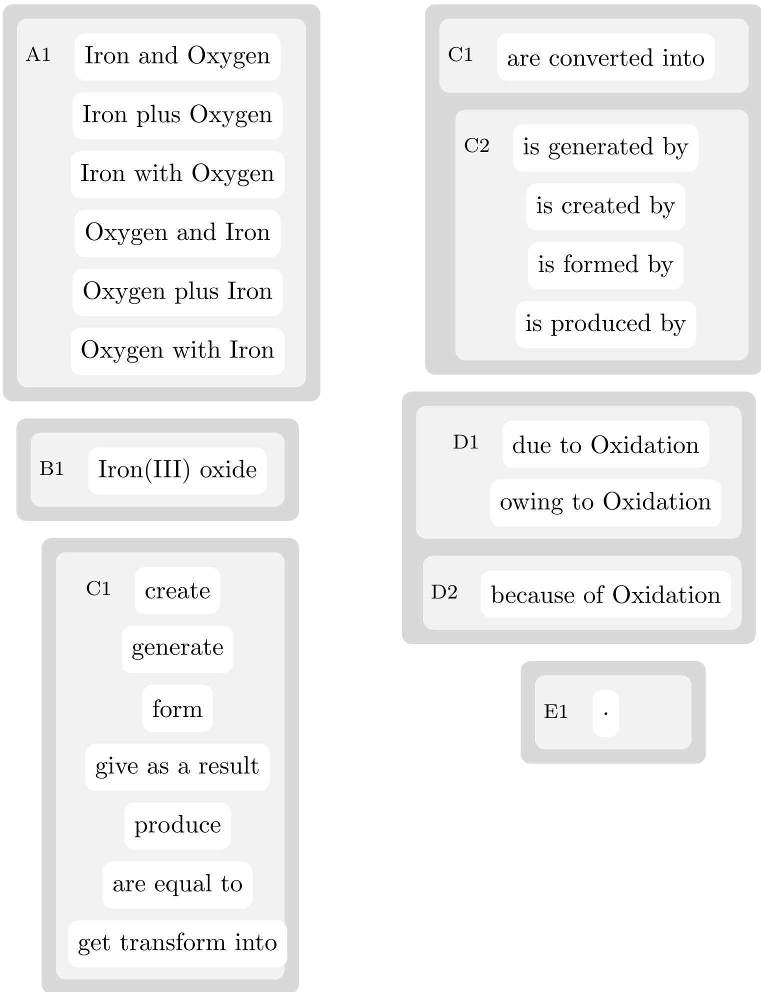
RCS 2 $< A, C1, B, D2, E >$

RCS 3 $< B, C2, A, D1, E >$

RCS 4 $< B, C2, A, D2, E >$

(b)

Figure H.27: ReSS specification RC and RCS for question 6 in activity ChemicalReactions-v1 by T1.



(a)

RCS 1 $< A, C1, B, D1, E >$

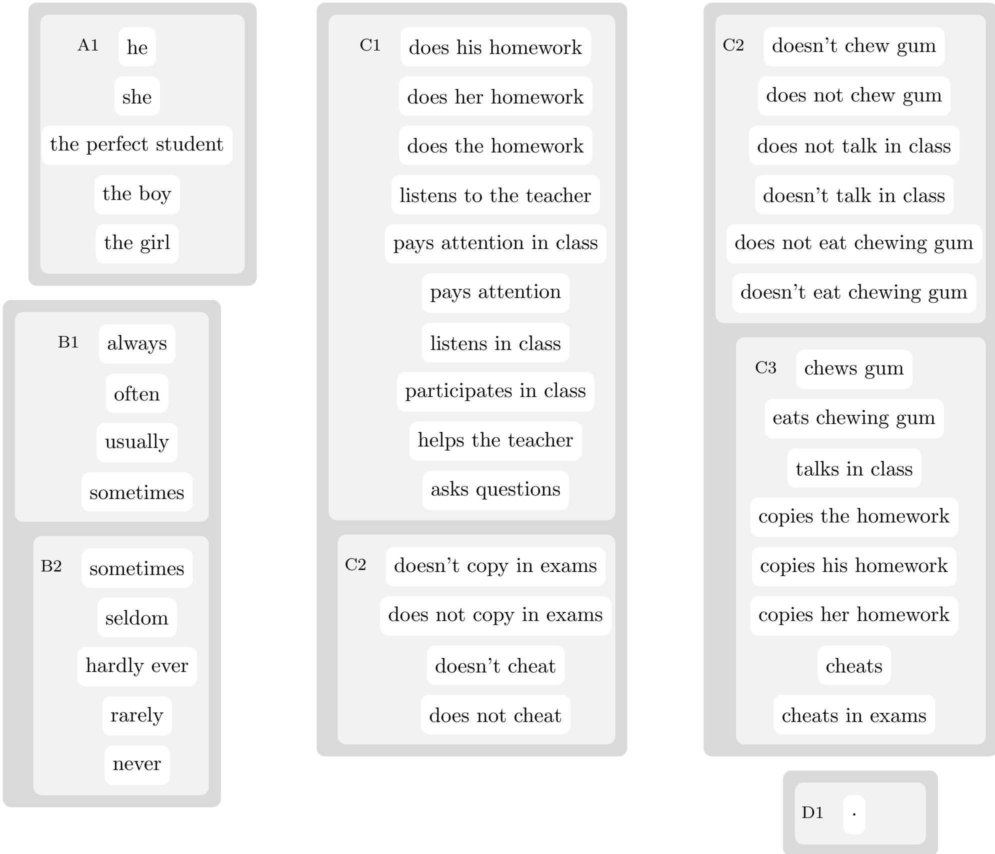
RCS 2 $< A, C1, B, D2, E >$

RCS 3 $< B, C2, A, D1, E >$

RCS 4 $< B, C2, A, D2, E >$

(b)

Figure H.28: ReSS specification RC and RCS for question 7 in activity ChemicalReactions-v1 by T1.



(a)

RCS 1 $< A, B1, C1, D >$

RCS 2 $< A, B2, C3, D >$

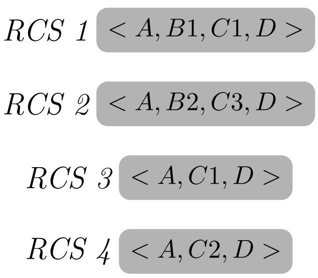
RCS 3 $< A, C1, D >$

RCS 4 $< A, C2, D >$

(b)

Figure H.29: ReSS specification RC and RCS for question 1 in activity PerfectStudent-v1 by T2/T3.

(a)



(b)

Figure H.30: ReSS specification RC and RCS for question 2 in activity PerfectStudent-v1 by T2/T3.

(a)

RCS 1 $< A, B1, C3, D >$

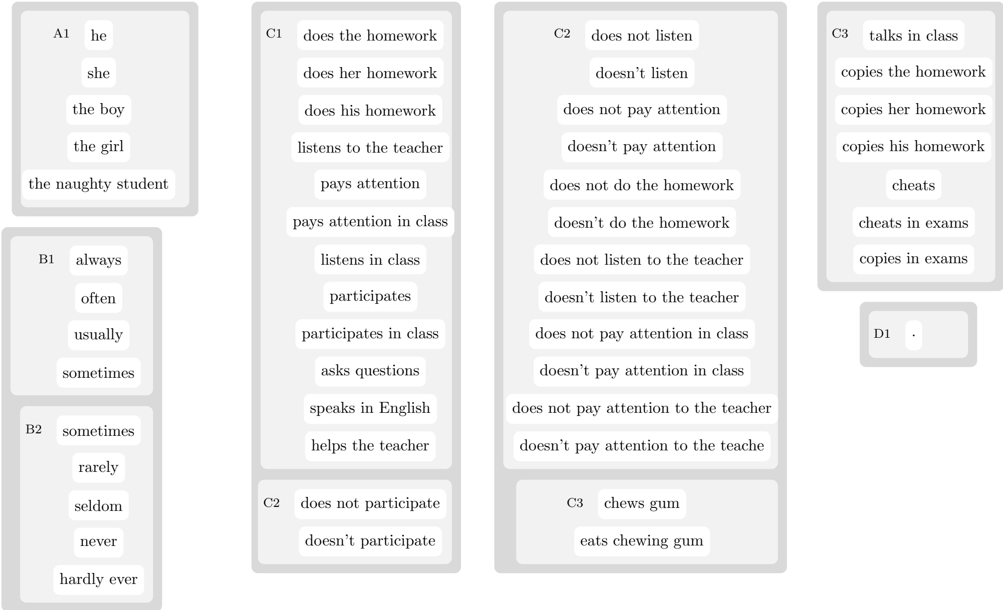
RCS 2 $< A, B2, C1, D >$

RCS 3 $< A, C2, D >$

RCS 4 $< A, C3, D >$

(b)

Figure H.31: ReSS specification RC and RCS for question 3 in activity PerfectStudent-v1 by T2/T3.



(a)

RCS 1 $< A, B1, C3, D >$

RCS 2 $< A, B2, C1, D >$

RCS 3 $< A, C2, D >$

RCS 4 $< A, C3, D >$

(b)

Figure H.32: ReSS specification RC and RCS for question 4 in activity PerfectStudent-v1 by T2/T3.

(a)

RCS 1	$< A, B1, G, E1, F, C >$	*RCS 10*	$< A, D, B1, G, E3, F, C >$
RCS 2	$< A, B1, G, E3, F, C >$	*RCS 11*	$< A, D, B1, G, E4, F, C >$
RCS 3	$< A, B1, G, F, E2, C >$	*RCS 12*	$< A, D, B1, G, F, C >$
RCS 4	$< A, B2, C >$	*RCS 13*	$< A, D, B1, G, F, E2, C >$
RCS 5	$< A, B2, G, E1, F, C >$	*RCS 14*	$< A, D, B2, G, E1, F, C >$
RCS 6	$< A, B2, G, E3, F, C >$	*RCS 15*	$< A, D, B2, G, E1, F, C >$
RCS 7	$< A, B2, G, F, E2, C >$	*RCS 16*	$< A, D, B2, G, E3, F, C >$
RCS 8	$< A, D, B1, C >$	*RCS 17*	$< A, D, B2, G, E4, F, C >$
RCS 9	$< A, D, B1, G, E1, F, C >$	*RCS 18*	$< A, D, B2, G, F, E2, C >$

(b)

Figure H.33: ReSS specification RC and RCS for question 1 in activity Routines1 by T2/T3.

(a)

RCS 1 $< A, B1, G, E1, F, C >$

RCS 2 $< A, B1, G, E3, F, C >$

RCS 3 $< A, B1, G, F, E2, C >$

RCS 4 $< A, B2, C >$

RCS 5 $< A, B2, G, E1, F, C >$

RCS 6 $< A, B2, G, E3, F, C >$

RCS 7 $< A, B2, G, F, E2, C >$

RCS 8 $< A, D, B1, C >$

RCS 9 $< A, D, B1, G, E1, F, C >$

RCS 10 $< A, D, B1, G, E3, F, C >$

RCS 11 $< A, D, B1, G, E4, F, C >$

RCS 12 $< A, D, B1, G, F, C >$

RCS 13 $< A, D, B1, G, F, E2, C >$

RCS 14 $< A, D, B2, G, E1, F, C >$

RCS 15 $< A, D, B2, G, E1, F, C >$

RCS 16 $< A, D, B2, G, E3, F, C >$

RCS 17 $< A, D, B2, G, E4, F, C >$

RCS 18 $< A, D, B2, G, F, E2, C >$

(b)

Figure H.34: ReSS specification RC and RCS for question 4 in activity Routines1 by T2/T3.

(a)

RCS 1 $< A, B1, G, E1, F, C >$

RCS 2 $< A, B1, G, E3, F, C >$

RCS 3 $< A, B1, G, F, E2, C >$

RCS 4 $< A, B2, C >$

RCS 5 $< A, B2, G, E1, F, C >$

RCS 6 $< A, B2, G, E3, F, C >$

RCS 7 $< A, B2, G, F, E2, C >$

RCS 8 $< A, D, B1, C >$

RCS 9 $< A, D, B1, G, E1, F, C >$

RCS 10 $< A, D, B1, G, E3, F, C >$

RCS 11 $< A, D, B1, G, E4, F, C >$

RCS 12 $< A, D, B1, G, F, C >$

RCS 13 $< A, D, B1, G, F, E2, C >$

RCS 14 $< A, D, B2, G, E1, F, C >$

RCS 15 $< A, D, B2, G, E1, F, C >$

RCS 16 $< A, D, B2, G, E3, F, C >$

RCS 17 $< A, D, B2, G, E4, F, C >$

RCS 18 $< A, D, B2, G, F, E2, C >$

(b)

Figure H.35: ReSS specification RC and RCS for question 7 in activity Routines1 by T2/T3.

H.8 Complexity of ReSS specifications

Item	RC	Var	Str	Sent
A1-T1				
1	5	7	38	3240
2	5	7	29	2808
3	5	7	28	1296
4	5	7	30	1728
5	5	7	31	2592
6	5	7	29	624
7	5	7	23	216
A2-T1				
1	6	11	25	236
2	6	11	25	236
3	6	8	34	768
4	6	11	27	236
5	6	11	27	236
6	6	11	29	720
7	6	11	27	312
8	6	11	27	236
A3-T1				
1	7	9	27	308
2	7	9	24	168
3	8	14	37	320
4	7	13	38	488
5	8	14	40	288
6	8	10	24	1080
7	10	11	29	832
8	7	10	29	480
A1-T2/3				
1–3	7	11	60	21736
4–6	7	11	50	11246
7–8	7	11	48	6632
A2-T2/3				
1	4	7	43	500
2	4	7	44	525
3	4	7	50	595
4	4	7	50	595

Table H.11: Total number of Response Components, Variants and Strings per question in the ICALL activities authored by teachers.

Bibliography

Bas Aarts and April McMahon, editors. *The handbook of English linguistics*. Blackwell, Oxford, 2006. ISBN: 1405113820. (Cited on page 235).

Steven Abney. Parsing by chunks. In Robert Berwick, Steven Abney, and Carol Tenny, editors, *Principle-Based Parsing*, chapter Parsing By Chunks. Kluwer Academic Publishers, Dordrecht, 1991. (Cited on page 41).

Steven Abney. Partial parsing via finite-state cascades. In *The Robust Parsing Workshop of the European Summer School in Logic, Language and Information (ESSLLI '96)*, pages 1–8, Prague, Czech Republic, 1996. URL http://www.vinartus.net/spa/97a.pdf. (Cited on pages 41 and 42).

Lourdes Aguilar, Àlex Alsina, Anna-Belén Avilés, Toni Badia, Sergio Balari, Gemma Boleda, Stefan Bott, Jenny Brumme, Carme Colominas, Anna Espunya, Josep Fontana, Jordi Fontseca, Àngel Gil, Carmen Hernández, Laia Mayol, Louise McNally, Carme de la Mota, Martí Quixal, Yolanda Rodríguez, Oriol Valentín, Enric Vallduví, and Teresa Vallverdú. Prado: Preparación automatizada de documentos. Technical report, Ministerio de Industria, Tecnología e Innovación (TIC 2000-1681), 2004. (Cited on page 96).

Alex Alsina, Toni Badia, Gemma Boleda, Stefan Bott, Àngel Gil, Martí Quixal, and Oriol Valentín. CATCG: a general purpose parsing tool applied. In *Proceedings of Third International Conference on Language Resources and Evaluation*, Las Palmas, Spain, 2002. URL http://mutis.upf.es/glicom/Papers/lrec2002/badia.pdf. (Cited on pages 99 and 101).

Luiz Amaral. *Designing Intelligent Language Tutoring Systems: integrating Natural Language Processing technology into foreign language teaching*. PhD thesis, The Ohio State University, 2007. (Cited on pages xxvii, 9, 10, 18, 20, 22, 23, 24, and 48).

Luiz Amaral and Detmar Meurers. Where does icall fit into foreign language teaching? 23rd Annual Conference of the Computer Assisted Language Instruction Consortium (CALICO), May 19, 2006. University of Hawaii, 2006. URL http://purl.org/net/icall/handouts/calico06-amaral-meurers.pdf. (Cited on page 22).

Luiz Amaral and Detmar Meurers. From recording linguistic competence to supporting inferences about language acquisition in context: Extending the conceptualization of student models for intelligent computer-assisted language learning. *Computer-Assisted Language Learning*, 21(4):323–338, 2008. URL http://purl.org/dm/papers/amaral-meurers-call08.html. (Cited on pages 19 and 23).

Luiz Amaral and Detmar Meurers. On Using Intelligent Computer-Assisted Language Learning in Real-Life Foreign Language Teaching and Learning. *ReCALL*, 23 (1):4–24, January 2011. URL http://purl.org/dm/papers/amaral-meurers-11.html. (Cited on pages 9, 10, 18, 19, 23, 24, 26, 27, 32, 33, and 43).

Luiz Amaral, Detmar Meurers, and Ramon Ziai. Analyzing learner language: Towards a flexible NLP architecture for intelligent language tutors. *Computer-Assisted Language Learning*, 24(1):1–16, 2011. URL http://purl.org/dm/papers/amaral-meurers-ziai-10.html. (Cited on pages 18, 22, and 24).

G. Antoniadis, S. Echinard, O. Kraif, T. Lebarbé, M. Loiseau, and C. Ponton. NLP-based scripting for CALL activities. In Lothar Lemnitzer, Detmar Meurers, and Erhard Hinrichs, editors, *Proceedings of eLearning for Computational Linguistics and Computational Linguistics for eLearning, International Workshop in Association with COLING 2004.*, pages 18–25, Geneva, Switzerland, August 28 2004. COLING. URL http://aclweb.org/anthology/W04-1703. (Cited on pages 19, 27, and 30).

Stewart Arneil and Martin Holmes. Juggling hot potatoes: decisions and compromises in creating authoring tools for the web. *ReCALL*, 11(2):12–19, 1999. doi: 10.1017/S0958344000004912. (Cited on pages 29 and 117).

Lyle F. Bachman. *Fundamental Considerations in Language Testing.* Oxford University Press, Oxford, UK, 1990. (Cited on page 61).

Lyle F. Bachman and Adrian S. Palmer. *Language Testing in Practice: Designing and Developing Useful Language Tests.* Oxford University Press, 1996. (Cited on pages 61, 62, 68, 105, 106, 117, 118, 120, 153, 198, and 199).

Toni Badia, Àngels Egea, and Antoni Tuells. Catmorf: multi two-level steps for catalan morphology. In *Proceedings of the Fifth conference on Applied Natural Language Processing*, pages 25–26, Morristown, NJ, USA, 1997. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/974281.974297. URL http://aclweb.org/anthology/A97-2015.pdf. (Cited on page 99).

Toni Badia, Gemma Boleda, Martí Quixal, and Eva Bofias. A modular architecture for the processing of free text. In *Workshop on Modular Programming applied to Natural Language Processing. EUROLAN 2001*, Iasi, Romania, July 2001. (Cited on pages 96, 99, and 101).

Toni Badia, Angel Gil, Marti Quixal, and Oriol Valentin. Nlp-enhanced error checking for catalan unrestricted text. In *Proceedings of Fourth International Conference on Language Resources and Evaluation*, volume VI, pages 1919–1922, Lisbon, Portugal, 2004. URL http://www.sdjt.si/bib/lrec04/pdf/260.pdf. (Cited on pages 96, 99, and 101).

Toni Badia, Lourdes Díaz, Sandrine Garnier, Rosa Lucha, Araceli Martinez, Martí Quixal, Ana Ruggia, and Paul Schmidt. ALLES correction tools: report on the quantitative and qualitative correction methodologies. Technical report, ALLES Project, 5th Framework Programme, 2005. (Cited on page 90).

Stacey Bailey and Detmar Meurers. Diagnosing meaning errors in short answers to reading comprehension questions. In Joel Tetreault, Jill Burstein, and Rachele De Felice, editors, *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications (BEA-3) at ACL'08*, pages 107–115, Columbus, Ohio, 2008. URL http://aclweb.org/anthology/W08-0913. (Cited on pages xxvii, 27, 28, 33, 150, 153, 198, and 221).

Stacey Bailey and Detmar Meurers. Exploring content assessment for ICALL. *CALICO Journal*, 2009. submitted. (Cited on pages 9, 28, 29, 68, 222, and 315).

Alan Bailin. Skills-in-context and student modeling. *CALICO Journal*, 8(1), September 1990. URL http://md1.csa.com/partners/viewrecord.php?requester=gs&collection=TRD&recid=13972116CI&q=Skills-in-Context+and+Student+Modeling&uid=790688827&setcookie=yes. (Cited on page 17).

Roberto Basili and Fabio Massimo Zanzotto. Parsing engineering and empirical robustness. *Nat. Lang. Eng.*, 8:97–120, June 2002. ISSN 1351-3249. doi: 10.1017/S1351324902002875. URL http://dl.acm.org/citation.cfm?id=973880.973882. (Cited on page 92).

Douglas Biber. Using register-diversified corpora for general language studies. *Computational Linguistics*, 19(2):219–241, 1993. (Cited on page 43).

Johnny Bigert and Ola Knutsson. Robust error detection: A hybrid approach combining unsupervised error detection and linguistic knowledge. In *Robust Methods in Analysis of Natural Language Data*, pages 10–19, Frascati, Italy, July 2002. URL http://www.johnnybigert.se/docs/BigertKnutsson02hybrid.pdf. (Cited on page 48).

Haji Binali, Vidyasagar Potdar, and Chen Wu. A state of the art opinion mining and its application domains. In *Proceedings of the IEEE International Conference on Industrial Technology (ICIT 2009)*, pages 1–6, Gippsland, Australia, 2009. IEEE. doi: 10.1109/ICIT.2009.4939640. (Cited on page 43).

John Bitchener, Stuart Young, and Denise Cameron. The effect of different types of corrective feedback on esl student writing. *Journal of Second Language Writing*, 14:191 – 205, 2005. doi: 10.1016/j.jslw.2005.08.001. (Cited on page 60).

Gemma Boleda. *Automatic acquisition of semantic classes for adjectives*. PhD thesis, Universitat Pompeu Fabra, 2007. (Cited on page 96).

Lars Borin. What have you done for me lately? The fickle alignment of NLP and CALL. Technical report, PLeaSe - PALaTe research report # 02, 2002. Presented at the EuroCALL 2002 pre-conference workshop on NLP in CALL, August 14, 2002, Jyväskylä, Finland. (Cited on page 37).

Nadjet Bouayad-Agha, Angel Gil, Oriol Valentin, and Victor Pascual. A sentence compression module for machine-assisted subtitling. In *CICLing*, pages 490–501, 2006. doi: 10.1007/11671299_51. (Cited on page 96).

Jose Roberto Boullosa, Martí Quixal, Paul Schmidt, José F. Esteban, and Angel Gil. Design of the overall architecture. Technical report, ALLES Project (IST-2001-34246), 2005. (Cited on pages 103 and 173).

Michael P. Breen and Christopher N. Candlin. The essentials of a communicative curriculum in language teaching. *Applied Linguistics*, I(2):89–112, 1980. URL http://applij.oxfordjournals.org/content/I/2/89.short. (Cited on page 52).

H. Douglas Brown. *Principles Of Language Learning and Teaching*. Pearson Education, 5th edition, 2007. (Cited on pages 6, 45, 52, 53, 54, 60, 63, 64, 65, and 68).

Ewa Buczowska and Richard M. Weist. The effects of formal instruction on the second-language acquisition of temporal location. *Language Learning*, 41(4):535–554, December 1991. doi: http://dx.doi.org/10.1111/j.1467-1770.1991.tb00692.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1467-1770.1991.tb00692.x/pdf. (Cited on page 6).

Susan Bull, Paul Brna, and Helen Pain. Extending the scope of the student model. *User Modeling and User-Adapted interaction*, 5:45–65, 1995. URL http://www.eee.bham.ac.uk/bull/papers-pdf/UMUAI-95.pdf. (Cited on page 19).

Hugh L. Burns and Charles G. Capps. Foundations of intelligent tutoring systems: An introduction. In M. Polson and J. J. Richardson, editors, *Foundations of Intelligent Tutoring Systems*, pages 1–20. Lawrence Erlbaum Associates Assosiates Publishers, 1988. (Cited on page 20).

Jill Burstein, Martin Chodorow, and Claudia Leacock. Criterion: Online essay evaluation: An application for automated evaluation of student essays. In *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-03)*, pages 3–10, Acapulco, Mexico, August 2003. URL http://ftp.ets.org/pub/res/erater_iaai03_burstein.pdf. (Cited on page 48).

Michael Carl and Antje Schmidt-Wigger. Shallow postmorphological processing with kurd. In *Proceedings of NeMLaP3/CoNLL98*, pages 257–265, 1998. (Cited on pages 95, 98, and 99).

Michael Carl, Johann Haller, Christoph Horschmann, and Axel Theofilidis. A hybrid example-based approach for detecting terminological variants in documents and lists of terms. In *6. Konferenz zur Verarbeitung natürlicher Sprache, KONVENS*, Saarbrücken, 2002. (Cited on page 96).

Stefano A. Cerri. ALICE: Acquisition of linguistic items in the context of examples. *Instructional Science*, 18:63–92, 1989. ISSN 0020-4277. doi: 10.1007/BF00121220. 10.1007/BF00121220. (Cited on page 17).

John Chandioux. 10 ans de météo. In *Traduction Assistée par Ordinateur. Actes du séminaire international sur la TAO et dossiers complémentaires.*, pages 169–173. Observatoire des Industries de la Langue (OFIL), Paris, 1989. Abbou, André (ed.). (Cited on page 7).

Jean Chandler. The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of l2 student writing. *Journal of Second Language Writing*, 12(3):267–296, 2003. ISSN 1060-3743. doi: 10.1016/S1060-3743(03)00038-9. URL http://www.sciencedirect.com/science/article/B6W5F-49D6T3P-1/2/c725c8890abe3b0efc29504cc1680fbb. (Cited on page 60).

Carol Chapelle. Multimedia call: Lessons to be learned from research on instructed sla. *Language Learning & Technology*, 2(1):21–39, July 1998. URL http://llt.msu.edu/vol2num1/pdf/article1.pdf. (Cited on page 275).

Liang Chen and Naoyuki Tokuda. A new template-enhanced ICALL system for a second language composition course. *CALICO Journal*, 20(3):561–578, 2003. URL https://www.calico.org/a-295-A%20New%20TemplateTemplateenhanced%20ICALL%20System%20for%20a%20Second%20Language%20Composition%20Course.html. (Cited on page 29).

Martin Chodorow and Claudia Leacock. An unsupervised method for detecting grammatical errors. In *Proceedings of the 1st annual meeting of the North American chapter of the Association for Computational Linguistics*, pages 140–147, Seattle, Washington, 29 April – 4 May 2000. Morgan Kaufmann. URL http://dl.acm.org/citation.cfm?id=974305.974324. (Cited on page 48).

Martin Chodorow, Joel Tetreault, and Na-Rae Han. Detection of grammatical errors involving prepositions. In *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions*, pages 25–30, Prague, Czech Republic, June 2007. URL http://aclweb.org/anthology/W07-1604. (Cited on page 48).

Josef Colpaert. Toward an ontological approach in goal-oriented language courseware design and its implications for technology-independent content structuring. *Computer Assisted Language Learning*, 19(2):109–127, 2006. (Cited on pages 27, 329, and 333).

Jack G. Conrad and Frank Schilder. Opinion mining in legal blogs. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law (ICAIL07*, pages 231–236. ACM Press, 2007. (Cited on page 43).

S. Pit Corder. Error analysis. In *The Edinburgh Course in Applied Linguistics*, volume 3, chapter 5, pages 122–154. Oxford Universtiy Press, 1974. (Cited on page 316).

S. Pit Corder. *Error Analysis and Interlanguage.* Oxford University Press, 1981. (Cited on page 329).

Daniel Dahlmeier and Hwee Tou Ng. Domain adaptation for semantic role labeling in the biomedical domain. *Bioinformatics*, 26(8):1098–1104, 2010. doi: 10.1093/bioinformatics/btq075. URL http://bioinformatics.oxfordjournals.org/content/26/8/1098.abstract. (Cited on page 43).

Fred J. Damerau. A technique for computer detection and correction of errors. *Communications of the ACM*, 7:171–176, 1964. (Cited on page 199).

Hal Daumé III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal Of Artificial Intelligence Research*, 26:101–126, 2006. doi: 10.1613/jair.1872. (Cited on page 43).

Alan Davies and Catherine Elder, editors. *The handbook of applied linguistics*. Blackwell, Malden, MA, 2004. ISBN: 0631228993. (Cited on page 235).

Kees de Bot. The psycholinguistics of the output hypothesis. *Language Learning*, 46 (3):529–555, September 1996. doi: 10.1111/j.1467-1770.1996.tb01246.x. URL http://repository.ubn.ru.nl/bitstream/2066/29108/1/29108___.PDF. (Cited on page 60).

William DeSmedt. Herr kommissar: An icall conversation simulator for intermediate german. In Holland et al. (1995), pages 153–174. (Cited on page 18).

R. Di Donato, M. Clyde, and J Vansant. *Deutsch, Na Klar! An Introductory German Course*. McGraw Hill, Boston, 2004. (Cited on page 22).

L. Dini and G. Malnati. Weak constraints and preference rules. In P. Bennett and P. Paggio, editors, *Studies in Machine Translation and Natural Language Processing*, chapter Weak constraints and preference rules, pages 75–90. Luxembourg: Commission of the European Communitie, 1993. (Cited on page 46).

C. Doughty. Second language instruction does make a difference: Evidence from an empirical study of sl relativization. *Studies in Second Language Acquisition*, 13:431–469, 1991. doi: http://dx.doi.org/10.1017/S0272263100010287. URL http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=2548976. (Cited on page 6).

C. Doughty. Instructed sla: Constraints, compensation and enhancement. In C. Doughty and M. Long, editors, *The handbook of Second Language Acquisition*, pages 256–310. Blackwell Publishing, Malden, MA, 2003. (Cited on page 6).

D. Douglas. *Assessing Languages for Specific Purposes*. Cambridge University Press, Cambridge, 2000. (Cited on page 118).

Shona Douglas and Robert Dale. Towards robust patr. In *Proceedings of the 14th conference on Computational linguistics - Volume 2*, COLING '92, pages 468–474, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/992133.992143. URL http://dx.doi.org/10.3115/992133.992143. (Cited on page 46).

Heidi Dulay, Marina Burt, and Stephen Krashen. *Language Two*. Oxford University Press, New York, 1982. (Cited on page 241).

Patricia A. Dunkel, Patricia A. Dunkel, and Patricia A. Dunkel. Computer-assisted language learning and testing: Research issues and practice. In P. Dunkel, editor, *Computer-Assisted language Learning and Testing – Reserch Issues and Practice*. Newbury House, 1991. (Cited on page 6).

Alan F. Duval, Louise Miller DuVal, Klaus Müller, and Herbert F. Wiese. *Moderne Deutsche Sprachlehre*. Random House, 2nd edition, 1975. (Cited on page 16).

Lourdes Díaz and Ana Ruggia. Cómo evaluar textos de fines específicos con ayuda de recursos informáticos: nuevas tecnologías al servicio del feedback en ele. *redELE – revista electrónica de didáctica / español lengua extranjera*, page Not paginated, March 2004. URL http://www.sgci.mec.es/redele/revista;http://www.educacion.es/redele/revista/diaz_ruggia.shtml. (Cited on page 91).

Lourdes Díaz, Ana Ruggia, and Martí Quixal. Links Among The Learning Tools. Technical report, ALLES Project (IST–2001–34246), 2003a. (Cited on pages 85, 86, 89, and 337).

Lourdes Díaz, Ana Ruggia, Martí Quixal, Enrique Torrejón, Jorge Jiménez, Celia Rico, Sandrine Garnier, and Paul Schmidt. Annex to D1.2 (Links Among The Learning Tools): Specifications of each of the learning units. Technical report, ALLES project (IST–2001–34246), 2003b. (Cited on pages 86, 87, and 88).

Lourdes Díaz, Ana Ruggia, Martí Quixal, Enrique Torrejón, Jorge Jiménez, Celia Rico, Sandrine Garnier, and Paul Schmidt. Links Among The Learning Tools And The Linguistic Tools. Technical report, ALLES Project (IST-2001-34246), 2004. (Cited on page 89).

Anas Elghafari, Detmar Meurers, and Holger Wunsch. Exploring the data-driven prediction of prepositions in english. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 267–275, Beijing, China, 2010. URL http://aclweb.org/anthology/C10-2031. (Cited on page 48).

Rod Ellis. *Task-based Language Learning and Teaching*. Oxford University Press, Oxford, UK, 2003. (Cited on pages 9, 53, 54, 55, 56, 60, 61, 63, 68, 78, 89, 105, 106, and 117).

Rod Ellis. Instructed language learning and task-based teaching. In Eli Hinkel, editor, *Handbook of Research in Second Language Teaching and Learning*, pages 713–728. Routledge, Mahwah, NJ, 2005. (Cited on pages 6 and 52).

Sheila Estaire and Javier Zanón. *Planning classwork: A task-based approach*. Educational Language Teaching. MacMillan-Heinemann, Oxford, 1994. (Cited on pages xxvii, 54, 55, 56, 57, 58, 59, 68, 76, 83, 85, 86, 87, 88, 90, 92, 104, 106, 117, 118, 329, 333, and 337).

Mariona Estrada, Raquel Navarro-Prieto, and Martí Quixal. Combined evaluation of a virtual learning environment: use of qualitative methods and log interpretation to evaluate a computer mediated language course. In *EDULEARN09*, Barcelona (Spain), 6th-8th of July 2009. URL http://faculty.ksu.edu.sa/7338/pdf/985.pdf. (Cited on page 84).

Rachele De Felice and Stephen Pulman. Automatic detection of preposition errors in learner writing. *CALICO Journal*, 26(3):512–528, May 2009. URL https://www.calico.org/a-758-Automatic%20Detection%20of%20Preposition%20Errors%20in%20Learner%20Writing.html. (Cited on page 48).

Jennifer Foster. Treebanks gone bad: Parser evaluation and retraining using a treebank of ungrammatical sentences. *International Journal on Document Analysis and Recognition*, 10(3–4):129–145, December 2007. URL http://www.springerlink.com/content/8260q041416j2618/. (Cited on pages 44, 48, and 49).

W. Nelson Francis and Henry Kučera. *Frequency Analysis of English Usage*. Houghton Mifflin, Boston, MA, 1982. (Cited on page 43).

Michael Gamon, Jianfeng Gao, Chris Brockett, Alexander Klementiev, William Dolan, Dmitriy Belenko, and Lucy Vanderwende. Using contextual speller techniques and language modeling for ESL error correction. In *Proceedings of IJCNLP*, Hyderabad, India, 2008. URL http://www.mt-archive.info/IJCNLP-2008-Gamon.pdf. (Cited on page 48).

Michael Gamon, Claudia Leacock, Chris Brockett, William B. Dolan, Jianfeng Gao, Dmitriy Belenko, and Alexandre Klementiev. Using statistical techniques and web search to correct esl errors. *CALICO Journal*, 26(3):491–511, May 2009. URL http://purl.org/calico/Gamon.Leacock.ea-09.pdf. (Cited on page 17).

Johann Gamper and Judith Knapp. A review of intelligent call systems. *Computer Assisted Language Learning*, 15(4):329–342, 2002. ISSN 0958-8221. URL http://www.informaworld.com/10.1076/call.15.4.329.8270. (Cited on page 10).

Sandrine Garnier, Paul Schmidt, Toni Badia, and Martí Quixal. Morphosyntactic taggers: report on the developing of tools for the evaluation of written text. Technical report, ALLES project (IST–2001–34246), 2003a. (Cited on pages 96 and 99).

Sandrine Garnier, Paul Schmidt, Toni Badia, Àngel Gil, and Martí Quixal. Syntactic parsers. Technical report, ALLES Project (IST-2001-34246), 2003b. (Cited on page 99).

Nina Garrett. *Modern Media in Foreign Language Education: Theory and Implementation*, chapter A psycholinguistic perspective on grammar and CALL, pages 169–196. National Textbook Company, 1987. William Flint Smith (Ed.). (Cited on page 19).

Daniel Gildea. Corpus variation and parser performance. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 167–202, Pittsburgh, PA, 2001. (Cited on page 43).

Andrew R. Golding and Dan Roth. A winnow-based approach to context-sensitive spelling correction. *Machine Learning*, 34(1-3):107–130, February 1999. ISSN 0885-6125. doi: 10.1023/A:1007545901558. URL http://dx.doi.org/10.1023/A:1007545901558. (Cited on page 48).

Andrew R. Golding and Yves Schabes. Combining trigram-based and feature-based methods for context-sensitive spelling correction. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, ACL '96, pages 71–78, Stroudsburg, PA, USA, 1996. Association for Computational Linguistics. doi: 10.3115/981863.981873. URL http://dx.doi.org/10.3115/981863.981873. (Cited on page 48).

Sylviane Granger. Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal*, 20(3):465–480, 2003. URL http://purl.org/calico/Granger03.pdf. (Cited on page 18).

Johann Haller. Multilint - a technical documentation system with multilingual intelligence. In *ASLIB*, 1996. (Cited on page 96).

Johann Haller. Multidoc-authoring aids for multilingual technical documentation. In J. et al. Chabás, editor, *Proceedings of the First International Conference on Specialized Translation*, pages 143–147, Barcelona, Universitat Pompeu Fabra, 2001. (Cited on page 96).

Johann Haller, Michael Carl, Sandrine Garnier, and Brigitte Stroede. NLP tools for intelligent learner utterance evaluation. In Rodolfo Delmonte, editor, *InSTIL/ICALL 2004 Symposium on Computer Assisted Learning, NLP and speech technologies in advanced language learning systems*, Venice, Italy, 2004. International Speech Communication Association (ISCA). URL http://project.cgm.unive.it/html/events/ICALL2004/papers/031haller.pdf. (Cited on page 96).

Henry J. Hamburger and R. Hashim. Foreign language tutoring and learning environment. In *Intelligent Tutoring Systems for Foreign Language Learning*. Springer-Verlag, New York, 1992. M. Swartz and M. Yazdani (Eds.). (Cited on page 18).

Trude Heift. *Designed Intelligence: A Language Teacher Model*. PhD thesis, Simon Fraser University, 1998. URL http://www.sfu.ca/~heift/pubs.html. (Cited on pages 10 and 22).

Trude Heift. Error-specific and individualized feedback in a web-based language tutoring system: Do they read it? *ReCALL*, 13(2):129–142, 2001a. (Cited on pages 22, 65, and 67).

Trude Heift. Intelligent language tutoring systems for grammar practice. *Zeitschrift für Interkulturellen Fremdsprachenunterricht*, 6(2):1–15, 2001b. URL http://zif.spz.tu-darmstadt.de/jg-06-2/beitrag/heift2.htm. (Cited on pages 22, 23, and 69).

Trude Heift. Multiple learner errors and meaningful feedback: A challenge for icall systems. *CALICO Journal*, 20(3):533–548, 2003. URL http://purl.org/calico/heift-03.html. (Cited on pages 10, 22, 23, 24, 26, 47, and 48).

Trude Heift. Corrective feedback and learner uptake in call. *ReCALL*, 16(2):416–431, 2004. URL http://journals.cambridge.org/production/action/cjoGetFulltext?fulltextid=265118. (Cited on pages 22, 65, 67, 68, and 69).

Trude Heift. Inspectable learner reports for web-based language learning. *ReCALL*, 17(1):32–46, 2005. URL http://journals.cambridge.org/action/displayFulltext?type=1&fid=305549&jid=&volumeId=&issueId=01&aid=305548&bodyId=&membershipNumber=&societyETOCSession=. (Cited on page 10).

Trude Heift. Prompting in call: A longitudinal study of learner uptake. *Modern Language Journal*, 94(2):198–216, 2010a. doi: 10.1111/j.1540-4781.2010.01017.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1540-4781.2010.01017.x/abstract. (Cited on pages 26, 27, and 32).

Trude Heift. Developing an intelligent language tutor. *CALICO Journal*, 27(3):443–459, May 2010b. URL https://calico.org/a-811-Developing%20an%20Intelligent%20Language%20Tutor.html. (Cited on pages 18, 22, and 23).

Trude Heift and Devlan Nicholson. Web delivery of adaptive and interactive language tutoring. *International Journal of Artificial Intelligence in Education*, 12(4):310–325, 2001. URL http://aied.inf.ed.ac.uk/members01/archive/vol_12/heift/paper.pdf. (Cited on pages 22, 24, 25, 26, and 48).

Trude Heift and Mathias Schulze. *Errors and Intelligence in Computer-Assisted Language Learning: Parsers and Pedagogues*. Routledge, 2007. (Cited on pages 6, 7, 9, 10, 17, 18, 19, 20, 27, 30, 33, and 47).

V. Holland, J. Kaplan, and M. Sams, editors. *Intelligent Language Tutors. Theory Shaping Technology*. Lawrence Erlbaum Associates, Inc., New Jersey, 1995. (Cited on pages 434 and 444).

Pierre Isabelle. Machine translation at the taum group. In Margaret King, editor, *Proceedings of the 3rd Lugano Tutorial, Lugano, Switzerland, 2–7 April 1984*, chapter 15, pages 247–277. Edinburgh University Press, 1987. (Cited on page 7).

Carl James. *Errors in Language Learning and Use: Exploring Error Analysis*. Longman, London and New York, 1998. (Cited on page 329).

Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, Upper Saddle River, NJ, 2000. URL http://www.cs.colorado.edu/~martin/SLP/Updates/index.html. (Cited on page 42).

Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech*

Recognition. Prentice Hall, Upper Saddle River, NJ, 2nd edition edition, 2009. (Cited on pages 39, 40, 41, 42, and 96).

Fred Karlsson. Constraint grammar as a framework for parsing running text. In *Proceedings of the 13th Conference on Computational Linguistics (COLING), Volume 3*, pages 168–173, Helsinki, Finland, 1990. Association for Computational Linguistics. URL http://aclweb.org/anthology/C90-3030. (Cited on page 99).

Fred Karlsson, Atro Voutilainen, Juha Heikkilä, and Arto Anttila, editors. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Number 4 in Natural Language Processing. Mouton de Gruyter, Berlin and New York, 1995. (Cited on pages 96, 99, and 100).

Judith Knapp. *A new approach to CALL content authoring*. PhD thesis, Universität Hannover, 2004. (Cited on page 18).

Karen Kukich. Techniques for automatically correcting words in text. *ACM Computational Surveys*, 24:377–439, December 1992. ISSN 0360-0300. doi: http://doi.acm.org/10.1145/146370.146380. URL http://doi.acm.org/10.1145/146370.146380. (Cited on page 199).

Stanley C. Kwasny and Norman K. Sondheimer. Relaxation techniques for parsing grammatically ill-formed input in natural language understanding systems. *American Journal of Computational Linguistics*, 7(2):99–108, April–June 1981. URL http://purl.org/net/Kwasny.Sondheimer-81.pdf. (Cited on page 46).

Michael Levy. *Computer-Assisted Language Learning: Context and Conceptualization*. Oxford University Press, New York, 1997. (Cited on pages 6, 10, 19, 29, 30, 229, 317, and 330).

Michael Levy and Glenn Stockwell. *CALL Dimensions: Options and Issues in Computer-Assisted Language Learning*. Lowrence Erlbaum Associates, Publishers, New Jersey, 2006. URL http://books.google.com/books?id=ho223xGvDt0C&lpg=PR11&ots=gN9SpCwmXV&dq=CALL%20Dimensions%3A%20Options%20and%20Issues%20in%20Computer-Assisted%20Language%20Learning&lr&pg=PR11#v=onepage&q&f=false. (Cited on pages 6, 19, 30, 33, and 229).

Sébastien L'Haire. Vers un feedback plus intelligent. les enseignements du project freetext. In *Proceedings of the Journée d'étude de l'ATALA on NLP and Language Learning*, pages 1–12, 2004. URL http://www.u-grenoble3.fr/lidilem/talal/actes/JourneeTALAL-041022-lhaire.pdf. (Cited on page 18).

Sébastien L'Haire and Anne Vandeventer Faltin. Error diagnosis in the FreeText project. *CALICO Journal*, 20(3):481–495, 2003. URL https://www.calico.org/a-290-ErrorDiagnosisintheFreeTextProject.html. (Cited on page 18).

William Littlewood. The task-based approach: some questions and suggestions. *ELT Journal*, 58(4):319–326, October 2004. URL http://faculty.ksu.edu.sa/yousif/3421/thetask-basedapproach2004.pdf. (Cited on pages 9, 54, 55, 57, 68, 105, 106, 117, 190, 206, 213, and 220).

M.H. Long and C.J. Doughty. *The Handbook of Language Teaching*. Blackwell Handbooks in Linguistics. John Wiley & Sons, 2011. ISBN 9781444350029. URL http://books.google.com/books?id=opOGjLmxDqgC. (Cited on page 53).

Michael Long. A role for instruction in second language acquisition. In K. Hyltenstam and M. Pienemann, editors, *Modelling and Assessing Second Language Acquisition*, Multilingual Matters, pages 77–100. Clevedon Avon, 1985. (Cited on page 53).

John Lyons. *Linguistic semantics: an introduction*. Cambridge University Press, 2nd edition, 1995. ISBN 9780521438773. URL http://books.google.es/books?id=Na2g1ltaKuAC. Reprinted in 2002. (Cited on pages 235 and 236).

Heinz-Dieter Maas. Mpro - ein system zur analyse und synthese deutscher wörter. In Roland Hausser, editor, *Linguistische Verifikation, Sprache und Information*. Max Niemeyer Verlag, Tübingen, 1996. (Cited on pages 95, 96, and 99).

Laura Martín, Celia Rico, Ana Ruggia, Lourdes Díaz, Martí Quixal, Toni Badia, Paul Schmidt, Mike Sharwood Smith, Antonio Sánchez Valderabanos, and Araceli Martínez. Final report. Technical report, ALLES Project (IST-2001-34246), 2005. (Cited on page 84).

Clive Matthews. Going AI: Foundations of ICALL. *Computer Assisted Language Learning*, 5(1):13–31, 1992. doi: 10.1080/0958822920050103. (Cited on page 19).

L. Mayol, G. Boleda, and T. Badia. Automatic acquisition of syntactic verb classes with basic resources. *Language Resources and Evaluation*, 39:295–312, 2005. ISSN 1574-020X. URL http://dx.doi.org/10.1007/s10579-006-9000-x. 10.1007/s10579-006-9000-x. (Cited on page 96).

Wolfgang Menzel. Robust processing of natural language. In *Proceedings of the 19th Annual German Conference on Artificial Intelligence*, pages 19–34. Springer, 1995. URL http://books.google.com/books?id=37YGKIf5AawC&lpg=PP1&pg=PA19#v=onepage&q&f=false. (Cited on page 44).

Wolfgang Menzel and Ingo Schröder. Error diagnosis for language learning systems. *ReCALL*, Special ed.:20–30, 1999. URL http://nats-www.informatik.uni-hamburg.de/~ingo/papers/recall99.pdf.gz. (Cited on page 18).

Detmar Meurers, Ramon Ziai, Luiz Amaral, Adriane Boyd, Aleksandar Dimitrov, Vanessa Metcalf, and Niels Ott. Enhancing authentic web pages for language learners. In *Proceedings of the 5th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-5) at NAACL-HLT 2010*, pages 10–18, Los Angeles, 2010. URL http://aclweb.org/anthology/W10-1002.pdf. (Cited on page 31).

Lisa Michaud and Kathleen McCoy. Capturing the evolution of grammatical knowledge in a call system for deaf learners of english. *International Journal of Artificial Intelligence in Education*, 16, 2006. (Cited on page 17).

Teruko Mitamura, Eric H. Nyberg, and Jaime G. Carbonell. Automated corpus analysis and the acquisition of large, multi-lingual knowledge bases for mt. In *In 5th International Conference on Theoretical and Methodological Issues in Machine Translation*, 1993. (Cited on page 42).

Noriko Nagata. *A Study of the Effectiveness of Intelligent CALI as an Application of Natural Language Processing*. PhD thesis, University of Pittsburgh, 1992. (Cited on page 21).

Noriko Nagata. Intelligent computer feedback for second language instruction. *The Modern Language Journal*, 77(3):330–339, 1993. URL http://www.usfca.edu/japanese/MLJ93.pdf. (Cited on pages 19, 22, 65, and 69).

Noriko Nagata. An effective application of natural language processing in second language instruction. *CALICO Journal*, 13(1):47–67, 1995. URL http://www.usfca.edu/japanese/CALICO95.pdf. (Cited on pages 19, 21, 22, 23, 24, 25, 65, and 69).

Noriko Nagata. Computer vs. workbook instruction in second language acquistion. *CALICO Journal*, 14(1):53–75, 1996. URL http://www.usfca.edu/japanese/CALICO96.pdf. (Cited on page 19).

Noriko Nagata. The effectiveness of computer-assisted metalinguistic instruction: A case study in Japanese. *Foreign Language Annals*, 30(2):187–200, 1997a. doi: 10.1111/j.1944-9720.1997.tb02341.x. (Cited on page 10).

Noriko Nagata. An experimental comparison of deductive and inductive feedback generated by a simple parser. *System*, 25(4):515–534, 1997b. URL http://www.sciencedirect.com/science/article/B6VCH-3SWSGM1-4/2/06a1a9a34cc6d2d2ca9217588720783b. (Cited on pages 21, 22, 23, 24, 25, 30, 65, and 66).

Noriko Nagata. Input vs. output practice in educational software for second language acquisition. *Language Learning & Technology*, 1(2):23–40, January 1998. URL http://llt.msu.edu/vol1num2/pdf/article1.pdf. (Cited on page 22).

Noriko Nagata. Banzai: An application of natural language processing to web based language learning. *CALICO Journal*, 19(3):583–599, 2002. URL http://www.usfca.edu/japanese/CALICO02.pdf. (Cited on pages 10, 21, 22, 23, 24, 25, 26, and 48).

Noriko Nagata. *ROBO-SENSEI: Personal Japanese tutor*. Cheng and Tsui, Boston, MA, 2004. (Cited on pages 10 and 21).

Noriko Nagata. Robo-sensei's NLP-based error detection and feedback generation. *CALICO Journal*, 26(3):562–579, May 2009. URL https://www.calico.org/a-762-RoboSenseis%20NLPBased%20Error%20Detection%20and%20Feedback%20Generation.html. (Cited on pages 20, 25, and 48).

Noriko Nagata. Some design issues for an online japanese textbook. *CALICO Journal*, 27(3):460–476, May 2010. URL http://calico.org/html/article_812.pdf. (Cited on pages 18, 21, 22, 23, 26, 32, and 33).

G Nelson, J Ward, S Desch, and R Kaplov. Two new strategies for computer aided language instruction. *Foreign Language Annals*, 9(1):28–37, February 1976. (Cited on page 16).

John Nerbonne, Duco Dokter, and Petra Smit. Morphological processing and computer-assisted language learning. *Computer Assisted Language Learning*, 11 (5):543–559, 1998. URL http://urd.let.rug.nl/nerbonne/papers/call_fr.pdf. (Cited on page 18).

David Nunan. *Research Methods In Language Learning*. Cambridge Language Teaching Library. Cambridge University Press, 1992. (Cited on page 330).

David Nunan. *Task-Based Language Teaching. Cambridge*. Cambridge University Press, 2004. (Cited on pages 52, 53, and 55).

David Nunan and Clarice Lamb. *The self-directed teacher: managing the learning process*. Cambridge Language Education. Cambridge University Press, fourth edition, 1996. Jack C. Richards, ed. (Cited on page 55).

Council of Europe. *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press, Cambridge, 2001. URL http://www.coe.int/t/dg4/linguistic/CADRE_EN.asp. (Cited on page 85).

Lluís Padró. *A Hybrid Environment for Syntax-Semantic Tagging*. PhD thesis, Dep. Llenguatges i Sistemes Informàtics. Universitat PolitÚcnica de Catalunya, February 1998. (Cited on page 40).

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *IN PROCEEDINGS OF EMNLP*, pages 79–86, 2002. (Cited on page 43).

Kathleen Marshall Pederson. Research on call. In Smith W. F., editor, *Modern Media in Foreign Language Education: Theory and Implementation*, pages 583–598. National Textbook Company, Lincolnwood, Illinois, 1988. (Cited on page 6).

Ken Petersen. *Implicit Corrective Feedback in Computer-Guided Interaction: Does Mode Matter?* PhD thesis, Georgetown University, 2010. URL http://purl.org/net/Petersen-10.pdf. (Cited on pages 65 and 66).

Fieny Pijls, Walter Daelemans, and Gerard Kempen. Artificial intelligence tools for grammar and spelling instruction. *Instructional Science*, 16:319–336, 1987. ISSN 0020-4277. URL http://dx.doi.org/10.1007/BF00117750. 10.1007/BF00117750. (Cited on page 17).

Barbara Plank and Gertjan van Noord. Grammar-driven versus data-driven: Which parsing system is more affected by domain shifts? In *ACL workshop NLP and Linguistics: Finding the Common Ground*, Uppsala, Sweden, July 2010. (Cited on page 43).

Carl Pollard and Ivan A. Sag. *Head-Driven Phrase Structure Grammar*. The University of Chicago Press, 1994. (Cited on page 24).

Joan-Tomàs Pujolà. Did call feedback feed back? researching learners' use of feedback. *ReCALL*, 13(1):79–98, 2001. doi: 10.1017/S0958344001000817. (Cited on pages 19, 20, 65, 66, 67, 69, and 117).

Joan-Tomàs Pujolà. Calling for help: researching language learning strategies using help facilities in a web-based multimedia program. *ReCALL*, 14(2):235–262, November 2002. ISSN 0958-3440. doi: 10.1017/S0958344002000423. URL http://portal.acm.org/citation.cfm?id=984914.984918. (Cited on page 117).

Carmen Pérez-Vidal. The integration of content and language in the classroom: A European approach to education (the second time around). In E. Dafouz and M. C. Guerini, editors, *LIL Across Educational Levels*, pages 3–17. Richmond publishing, Madrid: Santillana., 2009. (Cited on page 286).

Martí Quixal, Toni Badia, Beto Boullosa, Lourdes Díaz, and Ana Ruggia. Strategies for the generation of individualised feedback in distance language learning. In *Proceedings of the Workshop on Language-Enabled Technology and Development and Evaluation of Robust Spoken Dialogue Systems of ECAI 2006*, Riva del Garda, Italy, September 2006. (Cited on page 43).

Martí Quixal, Susanne Preuß, Beto Boullosa, and David García-Narbona. AutoLearn's authoring tool: a piece of cake for teachers. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 19–27, Los Angeles, June 2010. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.170.6551&rep=rep1&type=pdf. (Cited on page 277).

Kenneth Reeder, Trude Heift, Jörg Roche, Shahbaz Tabyanian, Stephan Schlickau, and Peter Gölz. Evaluating new media in language development. *Zeitschrift für Interkulturellen Fremdsprachenunterricht*, 6(2), 2001. URL http://zif.spz.tu-darmstadt.de/jg-06-2/beitrag/reeder1.htm. (Cited on page 22).

Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, volume 1, pages 448–453, Montréal, Canada, 1995. URL http://citeseer.ist.psu.edu/resnik95using.html. (Cited on page 40).

Jack Richards and Theodore Rodgers. *Approaches and Methods in Language Teaching*. Cambridge University Press, second edition, 2001. (Cited on pages 9 and 52).

Stephen D. Richardson and Lisa C. Braden-Harder. The experience of developing a large-scale natural language text procfassing system: Critique. In *ANLP*, pages 195–202, 1988. URL http://aclweb.org/anthology/A88-1027.pdf. (Cited on page 46).

Anne Rimrott and Trude Heift. Evaluating automatic detection of misspellings in german. *Language Learning and Technology*, 12(3):73–92, October 2008. URL http://llt.msu.edu/vol12num3/rimrottheift.pdf. (Cited on page 17).

Peter Robinson, editor. *Second Language Task Complexity: Researching the Cognition Hypothesis on language learning and performance.* John Benjamins, 2011. (Cited on page 331).

Douglas Roland and Daniel Jurafsky. How verb subcategorization frequencies are affected by corpus choice. In *In Proc. of the 36th Annual Meeting of the ACL*, pages 1122–1128, 1998. (Cited on page 43).

T. Roosmaa and G. Prószéky. *Language Teaching and Language Technology*, chapter GLOSSER - Using language technology tools for reading texts in a foreign language, pages 101–107. Lisse: Swets & Zeitlinger, 1998. S. Jager and J. A. Nerbonne and A. van Essen (Eds.). (Cited on page 18).

Christoph Rösener. A linguistic intelligent system for technology enhanced learning in vocational training: the illu project. In *Learning in the Synergy of Multiple Disciplines. 4th European Conference on Technology Enhanced Learning, EC-TEL 2009 Nice, France, Sept. 29, Oct. 2*, volume 5794(XVIII) of *Lecture Notes in Computer Science. Programming and Software Engineering*, page 813. Springer, Berlin, 2009. (Cited on page 29).

Ronald Rosenfeld. A maximum entropy approach to adaptive statistical language modeling. *Computer, Speech and Language*, 10:187–228, 1996. (Cited on page 43).

Ruth H. Sanders. Software review: e-tutor. *CALICO Journal*, 29(3):580–587, May 2012. (Cited on page 22).

Ruth H. Sanders and Alton F. Sanders. History of an ai spy game: Spion. In Holland et al. (1995), pages 141–151. (Cited on page 18).

Sandra J. Savignon. Communicative competence: theory and classroom practice, 1983. Reading. (Cited on page 52).

Michael Schoelles and Henry Hamburger. Teacher-usable exercise design tools. In *Proceedings of the Third International Conference on Intelligent Tutoring Systems*, pages 102–110, London, UK, 1996. Springer-Verlag. ISBN 3-540-61327-7. URL http://portal.acm.org/citation.cfm?id=648028.744989. (Cited on pages 8 and 18).

Josh Schroeder. Experiments in domain adaptation for statistical machine translation. In *Prague, Czech Republic. Association for Computational Linguistics*, pages 224–227, 2007. (Cited on page 43).

Mathias Schulze. Teaching Grammar – Learning Grammar: Aspects of Second Language Acquisition. *Computer Assisted Language Learning*, 11(2):215–228, 1998. (Cited on page 18).

Mathias Schulze. From the Developer to the Learner: Computing Grammar - Learning Grammar. *Re*, 11:117–124, 1999. (Cited on page 18).

Mathias Schulze. *Textana — Grammar and Grammar Checking in Parser-Based CALL*. PhD thesis, University of Manchester, 2001. (Cited on page 18).

Mathias Schulze. Grammatical errors and feedback: Some theoretical insights. *CALICO Journal*, 20(3):437–450, 2003. URL https://www.calico.org/a-287-Grammatical%20Errors%20and%20Feedback%20Some%20Theoretical%20Insights.html. (Cited on page 18).

Mathias Schulze. AI in CALL: Artificially Inflated or Almost Imminent? *CALICO Journal*, 25(3):510–527, May 2008. URL https://calico.org/a-714-AI%20in%20CALLArtificially%20Inflated%20or%20Almost%20Imminent1.html. (Cited on pages 16, 19, 26, and 333).

Mathias Schulze. Taking ICALL to Task. In M. Thomas and H. Reinders, editors, *Task-Based Language Teaching and Technology*, pages 63–82. Continuum, 2010. (Cited on pages 6, 17, 18, 19, 26, 27, 32, 33, and 73).

Mathias Schulze and Marie-Josée Hamel. NLP in CALL. *ReCA*, 10(2):55–56, 1998. (Cited on page 18).

Holger Schwenk. Building a statistical machine translation system for french using the europarl corpus. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 189–192, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1626355.1626380. (Cited on page 42).

Michael Sharwood Smith. Input enhancement in instructed SLA: Theoretical bases. *Studies in Second Language Acquisition*, 15:165–179, 1993. (Cited on page 31).

Diane H. Sonnenwald. Scientific collaboration: A synthesis of challenges and strategies. In B. Cronin, editor, *Annual review of information science and technology*, volume 41, chapter 14, pages 643–681. Information Today, Medford, NJ, 2007. (Cited on page 3).

Nina Spada. Form-focused instruction and second language acquisition: A review of classroom and laboratory research. *Language Teaching*, 30(2):73–87, 1997. doi: 10.1017/S0261444800012799. URL http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=2772824&fulltextType=RV&fileId=S0261444800012799. (Cited on page 53).

Oliver Streiter and Antje Schmidt-Wigger. The integration of linguistic and domain specific knowledge: Cat2 within anthem. *MT-News International*, 11:15–23, 1995. (Cited on page 96).

M. Swain. The output hypothesis and beyond: Mediating acquisition through collaborative dialogue. In J. P. Lantolf, editor, *Sociocultural theory and second language learning*, pages 97–114. Oxford University Press, Oxford, 2000. (Cited on page 60).

M. Swain. The output hypothesis: Theory and research. In E. Hinkel, editor, *Handbook on research in second language teaching and learning*, pages 471–484. Lawrence Erlbaum Associates, Mahwah, NJ, 2005. (Cited on page 60).

Merrill Swain and Sharon Lapkin. Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied Linguistics*, 16:370–391, 1995. URL http://applij.oxfordjournals.org/cgi/reprint/16/3/371.pdf. (Cited on page 60).

Pasi Tapanainen and Atro Voutilainen. Tagging accurately - don't guess if you know. In *Proceedings of Applied Natural Language Processing*, 1994. (Cited on page 40).

Pius ten Hacken. Computer-assisted language learning and the revolution in computational linguistics. *Linguistik Online*, 17:not paginated, 2003. URL http://www.linguistik-online.de/17_03/tenHacken.html. (Cited on page 7).

Joel Tetreault and Martin Chodorow. The ups and downs of preposition error detection in esl writing. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-08)*, pages 865–872, Manchester, UK, 2008. Association for Computational Linguistics. URL http://www.ets.org/Media/Research/pdf/r3.pdf. (Cited on page 48).

Janine Toole and Trude Heift. Task-generator: A portable system for generating learning tasks for intelligent language tutoring systems. In *Proceedings of ED-MEDIA 02, World Conference on Educational Multimedia, Hypermedia and Telecommunications, Charlottesville, VA*, pages 1972–1978. AACE, 2002a. (Cited on pages 30 and 31).

Janine Toole and Trude Heift. The tutor assistant: An authoring tool for an intelligent language tutoring system. *Computer Assisted Language Learning*, 15(4): 373–386, 2002b. ISSN 0958-8221. URL http://www.informaworld.com/10.1076/call.15.4.373.8266. (Cited on pages 29, 30, and 31).

John Truscott. Evidence and conjecture on the effects of correction: A response to Chandler. *Journal of Second Language Writing*, 13:337–343, 2004. URL http://www.hss.nthu.edu.tw/~fl/faculty/John/Evidence%20and%20Conjecture%202004.pdf. (Cited on page 60).

Lonneke van der Plas, James Henderson, and Paola Merlo. Domain adaptation with artificial data for semantic parsing of speech. In *Proceedings of NAACL*, Boulder, CO, USA, 2009. (Cited on page 43).

Duden Verlag. Korrektor 7.0. CD-ROM, 2010. (Cited on page 96).

Atro Voutilainen and Lluís Padró. Developing a hybrid nlp parser. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 80–87, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc. doi: http://dx.doi.org/10.3115/974557.974570. URL http://portal.acm.org/citation.cfm?id=974570. (Cited on page 40).

Joachim Wagner, Jennifer Foster, and Josef van Genabith. Judging grammaticality: Experiments in sentence classification. *CALICO Journal*, 26(3), May 2009. Special Issue of the 2008 CALICO Workshop on Automatic Analysis of Learner Language. (Cited on pages 48 and 49).

Ralph M. Weischedel and John E. Black. Responding intelligently to unparsable inputs. *American Journal of Computational Linguistics*, 6(2):97–109, 1980. URL http://aclweb.org/anthology/J80-2003. (Cited on page 46).

Ralph M. Weischedel, W. M. Voge, and M. James. An artificial intelligence approach to language instruction. *Artificial Intelligence*, 10(3):225–240, November 1978. doi: http://dx.doi.org/10.1016/S0004-3702(78)80015-0. (Cited on pages 15, 16, 17, 21, 29, 43, and 229).

Joseph Weizenbaum. *Computer Power and Human Reason: From Judgment to Calculation*. W. H. Freeman & Co., New York, 1976. ISBN 0716704641. (Cited on page 18).

Jane Willis. *A Framework for Task-Based Learning*. Longman Addison-Wesley, 1996. (Cited on pages 56, 329, and 333).

Kate Wolf-Quintero, Shunji Inagaki, and Hae-Young Kim. *Second language development in writing*. University of Hawaii Press, Honololu, 1998. (Cited on page 91).

D. J. Wood, J. S. Bruner, and G. Ross. The role of tutoring in problem solving. *Journal of Child Psychiatry and Psychology*, 17(2):89–100, 1976. (Cited on page 63).

Peter Wood. Turning language learners into linguists? first experiences of learners with a new corpus-driven language learning tool. In *CALICO conference*, Arizona State University, Tempe, Arizona, 2009. (Cited on page 18).

Ramon Ziai. A flexible annotation-based architecture for intelligent language tutoring systems. Master's thesis, Universität Tübingen, Seminar für Sprachwissenschaft, April 2009. URL http://www.sfs.uni-tuebingen.de/~rziai/papers/Ziai-09.pdf. (Cited on pages 18, 24, 25, 26, and 48).

Michael Zock. Swim or sink: The problem of communicating thought. In M. L. Swartz and M. Yazdani, editors, *Intelligent tutoring systems for foreign language learning: The bridge to international communication*, pages 235–247. Springer Verlag, Berlin, 1992. (Cited on page 17).