"Is There Choice in Non-Native Voice?"
Linguistic Feature Engineering and a Variationist Perspective in Automatic Native Language Identification

D i s s e r t a t i o n
zur
Erlangung des akademischen Grades
Doktor der Philosophie
in der Philosophischen Fakultät

der Eberhard Karls Universität Tübingen

vorgelegt von

Serhiy Bich
(ehem. Bykh)

aus

Dresden,
Deutschland

2017

**Gedruckt mit Genehmigung der Philosophischen Fakultät
der Eberhard Karls Universität Tübingen**

Dekan: Prof. Dr. Jürgen Leonhardt

**Hauptberichterstatter: Prof. Dr. Detmar Meurers
Mitberichterstatter: Prof. Dr. Heinrich Weber**

Tag der mündlichen Prüfung: 12. Mai 2017

TOBIAS-lib, Tübingen

Für Nadja und Sophia

Abstract

Is it possible to infer the native language of an author from a non-native text? Can we perform this task fully automatically? The interest in answers to these questions led to the emergence of a research field called *Native Language Identification (NLI)* in the first decade of this century. The requirement to *automatically* identify a particular property based on some *language data* situates the task in the intersection between computer science and linguistics, or in the context of computational linguistics, which combines both disciplines.

This thesis targets several relevant research questions in the context of NLI. In particular, what is the role of surface features and more abstract linguistic cues? How to combine different sets of features, and how to optimize the resulting large models? Do the findings generalize across different data sets? Can we benefit from considering the task in the light of the language variation theory?

In order to approach these questions, we conduct a range of quantitative and qualitative explorations, employing different machine learning techniques. We show how linguistic insight can advance technology, and how technology can advance linguistic insight, constituting a fruitful and promising interplay.

Acknowledgements

I would like to thank all those who supported me during my PhD time, and thus contributed an important part to the completion of this thesis: Be it a fruitful and helpful discussion in an apparent impasse, when after days or weeks of implementations and calculations, the final results were just the opposite of what was expected; or be it the warm and encouraging words whenever the accuracy curve just stayed frozen or decreased after the next model "improvement", and the consumed CPU power just pushed the figures on the electricity bills and contributed to the global warming without showing any noticeable benefits (I'm really very sorry for this waste and pollution!); or be it the time spent with my family and friends, or the amusing lunch and coffee breaks with the colleagues, which reminded me that life is much more than just improving results.

Now let me turn to more detail. I would like to first thank my advisor Detmar Meurers for agreeing to supervise this thesis and for his great support at all times throughout my PhD. I am very grateful to him for providing me the opportunity to work as a researcher and lecturer at the University of Tübingen, which contributed greatly to my professional and personal growth. I learned a lot from him about research and academia in general, as well as about computational linguistics and related disciplines in particular. Moreover, our discussions extended my knowledge on such useful things as how to raise children and on life in general. His helpful advice, vivid comments and constructive criticism, as well as the highly positive way of thinking encouraged and motivated me throughout my PhD. I thank my second advisor, Heinrich Weber, for agreeing to supervise this thesis despite being retired. I am very grateful for the fruitful discussions and his great support. He was also the one who got me started on my way in academia by providing the opportunity to work as a tutor for one of his grammar courses in the

beginning of my study in Tübingen. I would also like to thank Harald Schweizer for many interesting discussions in his office and during lunches in our favourite Chinese restaurant, and for providing me the opportunity to work as a research assistant at the Department of Computer Science, which contributed to a further improvement of skills relevant to the context of this thesis. Heinrich Weber and Harald Schweizer were also the advisors of my Masters' thesis on authorship attribution using recurring n-grams, which constituted the methodological starting point for this PhD thesis. I also thank the LGFG committee at the University of Tübingen for granting me a fellowship in the beginning of my research. I thank Harald Baayen, Gerhard Jäger and Friedrich Hamm for being together with my advisors on the thesis committee, and I thank to the committee for the many interesting comments and the thrilling 1.5h discussion during my defence.

I also thank all ex-colleagues and students at the SfS (University of Tübingen), with whom I had the great pleasure to work. I am very grateful for having had the chance to share the office with Jochen Saile. We had lots of enlightening discussions on basically any possible topic, ranging from research, teaching and server usage to "Bienenstich" cake and farming in Iowa. I am especially grateful to him for helping me to stay focused on the thesis by asking (basically daily) the painful but necessary question "How many pages did you complete today?". I was also frequently asked the same question by my ex-colleague Sowmya Vajjala, for which I am also very grateful (Actually, I already tried to express my gratitude to her by having asked her just the same question, whenever I had the chance to do it). I always enjoyed our frequent discussions on topics such as NLP, NLI and vegan food (including the way to prepare potatoes as "Salzkartoffeln" in Germany, questioned by Sowmya). I am also especially grateful to her for helping me to survive my first conference trip (COLING in Mumbai). I thank Ramon Ziai for the encouraging hallway talks on topics related and unrelated to our theses, and for the delicious coffee ("flow enhancer"), which helped me to stay focused on the thesis on days where there was almost no hope to finish a chapter which had to be finished according to the schedule. I also thank Petra Augurzky, Lisa Becker, Adriane Boyd, Xiaobin Chen, Maria Chinkina, Christl Glauder, Simón Ruiz Hernández, Julia Krivanek, Kordula De Kuthy, Niels Ott, Martí Quixal, Björn Rudzewitz and Julia Svetashova for a fruitful and enjoyable

time at SfS. I also thank all students who participated in the courses I taught at the SfS for all the interesting questions, comments and discussions during the lessons, and for staying awake.

I would also like to thank Kos, Samos and Rinia – no, this is not about holidays spent on Greek islands. These are just the names of the servers I most frequently used for the experiments in the context of this thesis. Sorry for all the broken HDD's, thrashing and other digital pain I caused to you.

Now, let me turn to the last and most important part of these acknowledgements. I thank my parents-in-law Susanne and Alwin, who live nearby, for taking care of me and supporting me just as if I was their own child. I will never be able to repay this in an appropriate way. Thank you very much! I thank my parents Lessja and Pawel, who live abroad, for the continuous caring support in any possible way. I would not have been able to get to this point without them. Thank you very much for everything! I thank my brother Stephan, my sister Nastja and my grandmother Hannah for being there for me, despite living far away. Finally, my last but most important words of thanks go to my wife Nadja and my daughter Sophia, to whom I dedicate this work. Thank you for your indescribable support, and your endless love and endurance. Thank you for enriching my life every single day!

Publications

Parts of this thesis appeared in the following peer-reviewed publications:

1. Serhiy Bykh & Detmar Meurers (2016). Advancing Linguistic Features and Insights by Label-informed Feature Grouping: An Exploration in the Context of Native Language Identification. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*. Osaka, Japan, pp. 739–749.

2. Serhiy Bykh & Detmar Meurers (2014). Exploring Syntactic Features for Native Language Identification: A Variationist Perspective on Feature Encoding and Ensemble Optimization. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*. Dublin, Ireland, pp. 1962–1973.

3. Detmar Meurers, Julia Krivanek & Serhiy Bykh (2014). On the Automatic Analysis of Learner Corpora: Native Language Identification as Experimental Testbed of Language Modeling between Surface Features and Linguistic Abstraction. In *Diachrony and Synchrony in English Corpus Studies*. Frankfurt a. M.: Peter Lang, pp. 285–314.

4. Serhiy Bykh, Sowmya Vajjala, Julia Krivanek & Detmar Meurers (2013). Combining Shallow and Linguistically Motivated Features in Native Language Identification. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-8) at NAACL-HLT 2013*. Atlanta, GA, pp. 197–206.

5. Serhiy Bykh & Detmar Meurers (2012). Native Language Identification Using Recurring N-grams – Investigating Abstraction and Domain Dependence. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*. Mumbai, India, pp. 425–440.

Funding

My research has been funded through an *LGFG Fellowship* (Landesgraduiertenförderung), and a position as *wiss. Angestellter* in the group of Prof. Detmar Meurers at the Seminar für Sprachwissenschaft (SfS), University of Tübingen.

Contents

I Introduction and the Context of this Thesis **1**

1 Introduction **3**

1.1 Native Language Identification (NLI) 4
1.2 The Goal and the Contributions of the Thesis 5
1.3 Research Questions 6
1.4 Outline of the Thesis 6

2 Research Context **8**

2.1 Second Language Acquisition (SLA) 8
2.2 Variationist Sociolinguistics (VS) 10
2.3 Machine Learning (ML) 12

3 Related Work **15**

3.1 Early Work on NLI 15
3.1.1 NLI in the CL context 15
3.1.2 NLI in the SLA context 21
3.1.3 Summary 23
3.2 The Contribution of the First NLI Shared Task 25
3.2.1 Data 26
3.2.2 Tasks 26
3.2.3 Approaches and Results 27
3.2.4 Summary 31
3.3 Current Trends in NLI 31
3.3.1 Cross-corpus Evaluation 32
3.3.2 L2s Different from English 32

3.3.3 New Features and Techniques . . . 33
3.3.4 Qualitative Analysis . . . 35
3.3.5 Summary . . . 37
3.4 Summary . . . 37

II Broad Linguistic Feature Exploration 39

4 Introduction 41

5 Systematically Exploring Recurring N-grams as Features 42
5.1 Introduction . . . 42
5.2 Systematic Feature Exploration . . . 43
5.2.1 Data . . . 43
5.2.2 Features . . . 43
5.2.3 Tools . . . 46
5.2.4 Results . . . 47
5.3 Cross-corpus Generalizability of the Findings . . . 52
5.3.1 Data . . . 52
5.3.2 Results . . . 54
5.4 Further Exploring Linguistic Generalization . . . 58
5.5 Conclusions . . . 62

6 Exploring Linguistic Features and Feature Combinations 65
6.1 Introduction . . . 65
6.2 Tasks . . . 66
6.3 Data . . . 66
6.4 Features . . . 68
6.4.1 Recurring N-grams . . . 69
6.4.2 Dependency . . . 70
6.4.3 Constituency . . . 71
6.4.4 Morphology . . . 71
6.4.5 Complexity . . . 72
6.4.6 Other . . . 73

6.5 Evaluation Setup 74
6.6 Classifier Models 75
6.6.1 Single Features 75
6.6.2 Ensembles 75
6.7 Results 76
6.7.1 Single Features 76
6.7.2 Ensembles 78
6.8 Conclusions 85

III A Variationist Approach to NLI 89

7 Introduction 91
7.1 A Variationist Perspective on NLI 91
7.2 Implementing a Variationist Perspective: Core Questions 92
7.3 Relevant Related Work in NLI 93

8 Variationist Feature Engineering 98
8.1 Linguistic Variables Revisited 98
8.2 Types of Linguistic Variables 101
8.2.1 Relative vs. Absolute 102
8.2.2 Lexical vs. Grammatical 103
8.2.3 Level of Granularity 104
8.3 Label-informed Feature Grouping 105
8.3.1 Introduction 105
8.3.2 Hierarchical Clustering 107
8.3.3 Clustering Linguistic Variables 109
8.4 Conclusions 116

9 Quantitative Explorations of the Variationist Approach 118
9.1 Introduction 118
9.2 Syntactic Category Realization (CFGR) 118
9.2.1 Data 119
9.2.2 Tools 119

9.2.3 Features . . . 119
9.2.4 Results . . . 122
9.2.5 Conclusions . . . 126
9.3 Verb Subcategorization (Verb Subcat) . . . 126
9.3.1 Data . . . 127
9.3.2 Tools . . . 127
9.3.3 Features . . . 127
9.3.4 Results . . . 130
9.3.5 Conclusions . . . 143
9.4 Conclusions . . . 144

10 Qualitative Explorations of the Variationist Approach 147
10.1 Introduction . . . 147
10.2 Evaluation Setup . . . 148
10.2.1 Data . . . 148
10.2.2 Tools . . . 149
10.2.3 Feature Choice and Encoding . . . 149
10.2.4 Feature Evaluation . . . 150
10.2.5 Feature Grouping . . . 154
10.2.6 Settings . . . 155
10.3 Explorations . . . 157
10.3.1 Subject (Pronoun) Realization . . . 157
10.3.2 Nominal Modification . . . 171
10.4 Conclusions . . . 175

11 Advantages and Limitations of the Variationist Approach 177

IV Advancing Performance 181

12 Introduction 183

13 Ensemble Optimization and Tuning Approach 185
13.1 Generating Ensembles . . . 185
13.2 Ensemble Optimization . . . 186

13.3 Ensemble Tuning 187
13.4 Conclusions 188

14 Advancing Performance Using Ensembles 189
14.1 Data 189
14.2 Features 189
14.3 Tools 191
14.4 Results 191
14.4.1 CFGR 192
14.4.2 Recurring N-grams 194
14.4.3 Combining Recurring N-grams and CFGR 195
14.4.4 Maximizing Performance Using Linguistic Features . . . 197
14.5 Conclusions 201

V Conclusions 205

15 Summary and Outlook 207
15.1 Summary 207
15.1.1 Broad Linguistic Feature Exploration 208
15.1.2 A Variationist Approach to NLI 209
15.1.3 Advancing Performance 216
15.2 Contributions 218
15.3 Limitations and Outlook 223

Zusammenfassung 225

Bibliography 226

Appendices

A Analysing Subject (Pronoun) Realization: Underlying Data 249

B Analysing Nominal Modification: Underlying Data 263

Part I

Introduction and the Context of this Thesis

Chapter 1

Introduction

Language is a substantial and fascinating part of our cognition, one of the central devices defining us as human beings. Moreover, the manifold and flexible nature of language and the versatile ways of using language, makes it also a part of collective identity, and of personality in particular. This raises a range of interesting questions, which become increasingly relevant in different security and commercial settings nowadays, facilitated by the possibilities and capabilities emerging in the era of enormous computational power and "big data". Is it possible to identify the *author* of a anonymous text? (Mosteller & Wallace, 1964; Holmes, 1994; Hoover, 2002) Can we automatically infer the *age* or the *gender* of a writer? (Argamon et al., 2009; Estival et al., 2007) Is it possible to identify the *native language* of the author based on a non-native text production? By asking the last question, we already arrived at the heart of this study.

In particular, this work is an interdisciplinary thesis in *Computational Linguistics*, focused on a specific natural language processing task, namely, the *Native Language Identification*. It is situated in the intersection between *Computer Science / Machine Learning*, *Second Language Acquisition* and *Variationist Sociolinguistics*. The thesis targets *quantitative* as well as *qualitative* aspects of the given task. On the one hand, it shows how incorporating linguistic knowledge can improve the accuracy of native language identification systems. On the other hand, it makes explicit how the output of such systems can be used to further advance the linguistic insight.

1.1 Native Language Identification (NLI)

> *The Gileadites captured the fords of the Jordan leading to Ephraim, and whenever a survivor of Ephraim said, "Let me cross over", the men of Gilead asked him, "Are you an Ephraimite?" If he replied, "No", they said, "All right, say 'Shibboleth.'" If he said, "Sibboleth", because he could not pronounce the word correctly, they seized him and killed him at the fords of the Jordan.*
>
> Judges 12, 5-6

We start by defining the task in the focus of this thesis, namely the *Native Language Identification*:

> *(Automatic) Native Language Identification* (NLI): The task of automatically identifying the native language of an author based on texts written in a second language, or any other language, different from the native language.

The task of NLI is of interest and high relevance for different reasons. On the one hand, NLI can be used as a testbed for data-driven, empirical exploration and verification of different hypotheses regarding the existence and the nature of cross-linguistic influence (*L1-transfer*), which is of *conceptual* relevance in the context of the Second Language Acquisition research. On the other hand, NLI is of *practical* relevance for a range of applications in the author profiling and security settings, or for learner modelling in the context of intelligent language tutoring systems, etc. (Argamon et al., 2009; Amaral & Meurers, 2008; Bykh & Meurers, 2012; Estival et al., 2007; Malmasi & Dras, 2014a). Hence, NLI provides an opportunity to advance theoretical insights and to build useful or – depending of the context – even critical tools and applications. Fortunately, due to the increasing availability of language learner data and the rapid technical advance, the application of increasingly powerful statistical techniques becomes feasible (see Section 2.3), which makes building high-performing NLI systems viable.

1.2 The Goal and the Contributions of the Thesis

In this thesis, we contribute to the two important aspects of NLI – the quantitative and the qualitative.

On the one hand, we are interested in the *quantitative* aspect of the task, i.e., our goal is to advance the classification performance of NLI systems. In this regard, the contribution of this thesis is as follows:

- Our aim is to design feature sets, capable of capturing general distinctive cues in the language use of speakers with different native language backgrounds. We approach that problem by broad linguistic feature engineering at different levels of linguistic modelling. We propose a range of features, novel for the task of NLI, and explore their single- and cross-corpus performance.

- Given a range of different feature types, the question is how to combine them, and how to optimize the corresponding complex models? Here, we explore combining a range of features using ensemble classifiers, and propose a technique for optimizing and tuning them.

On the other hand, we are interested in the *qualitative* aspect. Our goal is to explore the benefits, NLI could provide for the SLA research. In that regard, the contribution of this thesis is as follows:

- We explore what insight can be transferred from the data-driven outcomes in the context of NLI to the SLA research. For this, we focus on a particular sort of linguistically-motivated features, namely, features following a variationist perspective, and – in referring to the question in the title of this thesis with the term *voice* in its metaphorical sense – we investigate the choices made by the writers with different native language backgrounds, when producing non-native texts.

In the following section, we outline the core research questions, which are in the focus of this thesis.

1.3 Research Questions

In this thesis we will target the following five research questions:

1. How useful are features on different levels of linguistic modelling for the specific task of NLI?
   [LINGUISTIC-FEATURES][1]

2. How well do results and findings based on a broad range of features generalize across different data sets?
   [CROSS-CORPUS]

3. How can we abstract over individual features to obtain insights into the general underlying linguistic structures reflected in NLI?
   [GENERAL-STRUCTURES]

4. Can the application of variationist perspective to language data enhance an NLI system and contribute relevant SLA insight?
   [VARIATIONIST-PERSPECTIVE]

5. How can we optimize large models incorporating a broad range of features?
   [MODEL-OPTIMIZATION]

1.4 Outline of the Thesis

The rest of this thesis is organized as follows:

The remainder of Part I, namely Chapter 2 and Chapter 3, further clarify the context of this thesis, and discuss the related work on NLI respectively.

Part II describes our results utilizing a broad range of surface-based and linguistic features. In particular, Chapter 4 presents a brief introduction and outline. Chapter 5 shows our findings employing recurring n-grams, and Chapter 6 is dedicated to exploring a broader features space.

Part III reports our findings on adapting and applying a particular linguistic theory, namely, a variationist sociolinguistics perspective to the task of NLI.

[1] In the course of this thesis, we will use the shorthand notations of the form "[…]", listed after each of the research questions, as reference to these.

In particular, Chapter 7 clarifies some core questions in connection with the approach, and discusses relevant related work. Chapter 8 discusses the central notion of a linguistic variable, and proposes a revised definition for it. Furthermore, it presents a taxonomy of linguistic variables, we consider useful for this study. Finally, it describes a technique for generating more abstract variables by feature grouping. Chapter 9 and Chapter 10 present the evaluation of our variationist approach, and report our quantitative and qualitative findings respectively.

Part IV is concerned with further advancing the performance of NLI systems. Especially, it targets the question of how to combine a range of features, and how the corresponding complex models can be optimized and tuned. In particular, Chapter 12 presents a brief introduction and outline. Chapter 13 presents our ensemble approach, while Chapter 14 implements it and reports our findings.

Finally, Part V, consisting of Chapter 15, summarizes our results, and the particular contributions with respect to the research questions targeted in this thesis, as well as discusses the limitations and potential extensions to our work.

Chapter 2

Research Context

This thesis is situated in the intersection of several disciplines, namely *Second Language Acquisition*, *Variationist Sociolinguistics* and *Computer Science / Machine Learning*. In the following sections, we introduce each of them and present the relevant terminology.

2.1 Second Language Acquisition (SLA)

From the linguistic perspective, this thesis is first of all situated in the context of the *Second Language Acquisition (SLA)* research (Doughty & Long, 2003a; Gass et al., 2013; Krashen, 1982), and is mainly concerned with the fundamental notion of *cross-linguistic influence* or *L1-transfer* (Odlin, 1989, 2003; Dechert & Raupach, 1989; Selinker, 1969; Ortega, 2009; Gass & Selinker, 1992; Gass et al., 2013; Lado, 1957; Weinreich, 1953). In the following, based on the work cited above, we introduce some basic SLA terminology relevant here.

- *Native Language (L1)*: The first language that a child learns. It is also known as the primary language, the mother tongue, or the L1.
- *Second Language (L2)*: In general, this refers to any language learned after the L1. It may mean the second, third, tenth, etc. language.
- *Target Language (TL)*: The language being learned.

- *Learners*: Individuals learning a L2.
- *Interlanguage (IL)*: The language system, constructed by each learner at any given point in development.
- *Second Language Acquisition (SLA)*: The process of learning another language after the L1 has been learned.
- *L1-transfer*: The influence resulting from similarities and differences between the target language and any other previously acquired language. It affects all linguistic levels, including pragmatics and rhetoric, semantics, syntax, morphology, phonology, phonetics, and orthography. The L1-transfer can be negative or positive:
  - *Negative*: It emerges if the L1 and L2 substantially differ, but the learners are still using some L1 items or applying some L1 rules in L2 communication, which usually results in errors.
  - *Positive*: It emerges if there is a substantial similarity between L1 and L2, and learners can reuse the items or rules from L1 in L2 communication, quickly resulting in correct (or acceptable) L2 production.

In general, the research on SLA aims at discovering the nature and the sources of the underlying L2 knowledge system, as well as on explaining developmental success and failure (Doughty & Long, 2003b). In this context, cross-linguistic influence seems to play a central role. In fact, the early research in the 1950s and 1960s focused on the *contrastive analysis* (Lado, 1957; Stockwell et al., 1965), aimed at explaining and predicting difficulties in acquiring L2, based on differences between L1 and L2, thus focusing on L1-transfer. Hence, analysing different language pairs and comparing linguistic systems was predominant at that time. However, there emerged evidence that this perspective might be too narrow: There were cases where cross-linguistic comparisons failed to predict the actual difficulties, and at the same time, some of the predicted difficulties did not emerge in the language productions (cf. Odlin, 1989, 2003; Gass, 1996). So, the research moved the focus from the contrastive to the *error analysis* (Corder, 1967; Richards, 1971, 1974, cf. also Long & Sato, 1984). The new direction did

not deny certain phenomena related to the cross-linguistic influence, especially to the negative L1-transfer, but it considered corresponding findings as related only to a certain type of errors – i.e., interlanguage or interference errors – as part of a broader taxonomy, including, e.g., intralingual and developmental errors, etc. (Corder, 1967; Richards, 1971; Wong & Dras, 2009). Recent research shifted the focus from considering errors in particular to a more general perspective. The idea is that learners with different L1s might differ in the usage frequencies of certain language units or structures, i.e., there might be significant overuse/underuse patterns in their productions, which are not necessarily erroneous (Granger et al., 2002b; Ortega, 2009; Lüdeling, 2011; Jarvis & Crossley, 2012). While the investigation of positive and negative L1-transfer has apparently lost its predominant position in the SLA research, the idea of cross-linguistic influence still remains important. In particular, it seems to get revived by the research on NLI – an NLP classification task, drawing in the first place on the L1-transfer idea as its theoretical background.

Connections to this thesis In this thesis, we explore possible L1-transfer effects via NLI, utilizing a broad range of surface-based and linguistically-motivated features compiled in a data-driven way. We show how NLI techniques can help discovering new instructive L1-transfer candidates, as well as formulating and testing new hypotheses about L1-transfer, thus advancing SLA insight.

2.2 Variationist Sociolinguistics (VS)

Variation is an inherent part of language (Labov, 1969; Tagliamonte, 2012). It can be observed essentially everywhere, "from a conversation you overhear on the street to a story you read in the newspaper" (Tagliamonte, 2012, p. 2). In particular, different forms can be used to express (more or less) the same meaning. The core question is how to explain such language choices? The desideratum can be defined as finding "the order, the system, in the variation chaos" (Tagliamonte, 2012, p. 2). In fact, the research tradition on language variation turns out to be very old. That phenomenon was observed already in ancient times by Aristotle (Aristotle, 1933). It was approached, e.g., in the context of research on lexical

and structural synonymy, concerned with investigating the similarity and usage context of related words and structures (cf., e.g., Stanojević, 2009; Wulff, 2006; Weber, 2012, 2014). Eventually, one of the most comprehensive approaches to the study of language variation was established by the field of *Variationist Sociolinguistics* (VS) (Labov, 1972; Tagliamonte, 2012; Oliva & Serrano, 2013; Geeslin & Long, 2014). In the following we discuss the core principles of the VS.

Every investigation in the context of VS begins with the isolation and definition of the so-called *linguistic variable*, which can be realized by a set of *variants*. In the following we define and describe these notions in more detail (based on Tagliamonte, 2012).

- *Linguistic Variable*: In its most basic definition, a linguistic variable is two or more ways of saying the same thing.
- *Variant (of a Linguistic Variable)*: A particular option or choice in the context of a linguistic variable. In the basic definition, the variants should be equivalent regarding the expressed meaning, and the choice of a particular variant in a particular context must be systematic.

In general, each linguistic variable must show the following properties:

1. (at least) two different ways of saying the same thing;
2. is an abstraction;
3. is made up of variants;
4. comprises a linguistically defined set of some type:
    - a phoneme
    - a lexical item
    - a structural category
    - a natural class of units
    - a syntactic relationship
    - the permutation or placement of items

5. the variants of the variable must have a structurally defined relationship in the grammar;

6. the variants of the variable must co-vary, correlating with patterns of social and/or linguistic phenomena.

As pointed out, some variants can have a social meaning. In that case, the corresponding linguistic variables are called *sociolinguistic variables*. For example, in William Labov's study "The Social Stratification of (r) in New York City Department Stores" Labov (1972), he found that the presence or absence of the consonant [r] in postvocalic position (e.g., *car*, *fourth*) correlates with the ranking of people in status or prestige (social stratification). In general, the preference for particular variants can be indicative for different individual characteristics, such as the proficiency or the L1 (Young, 1991; Callies & Szczesniak, 2008; Callies & Zaytseva, 2011; Lüdeling, 2011; Meurers et al., 2014; Bykh & Meurers, 2014).

In Chapter 8 we discuss the notion of the linguistic variable in more detail, and suggest a revised version of its definition, which we consider more suitable in the context of this thesis.

Connections to this thesis Isolating suitable linguistic variables and revealing potential connections between variant choices and the L1s of the learners in a data-driven way, constitutes one of the core aims in this thesis.

2.3 Machine Learning (ML)

The field of *Machine Learning* is nowadays one of the most vivid and important areas in Computer Science, represented, e.g., by a range of international top-level conferences such as ICML (*International Conference on Machine Learning*), KDD (*International Conference on Knowledge Discovery and Data Mining*) or NIPS (*Neural Information Processing Systems*).

> *Machine Learning* (ML): Essentially, ML is about techniques for finding patterns or regularities in data, with the aim to explain that data and make predictions from it (Witten et al., 2011; Alpaydin, 2004; Kubat, 2015).

The rapid developments in the field of ML are supported by the immense advances in hardware technology, and inexorable increase of available data sources and data volume, often referred to by "big data". That allows to employ increasingly complex and powerful algorithms to tackle many of important and critical real-world tasks. ML has been applied in a range of different areas, such as medicine and healthcare, manufacturing and market, networks and communication, and, among many others, in science in general and in Computational Linguistics (CL) in particular (Witten et al., 2011; Alpaydin, 2004; Jurafsky & Martin, 2009).

Regarding the CL context, it is practically impossible to process and try to understand the patterns behind the ever-increasing amount of language data available from different sources, first of all on the World Wide Web, without the application of any ML techniques. Most of the core applications in CL such as Part-of-Speech tagging, Speech Recognition or any sort of Document Classification, etc. require the usage of ML techniques in order to achieve a state-of-the-art performance (Manning & Schütze, 1999; Jurafsky & Martin, 2009; Dickinson et al., 2013).

There are two main branches in ML (Witten et al., 2011; Alpaydin, 2004):

- *Supervised Learning*: ML techniques based on labelled training data. "Labelled" means that the label, i.e., the output value (discrete or numeric) for each of the instances, we deal with (e.g., words or texts) is known in advance. That knowledge is used to *train* the system, i.e., to learn a mapping from the input to a particular output value. After the training is finished, the system can be used to make predictions for new instances.

  Example in CL: *Document Classification/Categorization* – automatically assigning some label of interest to a document, e.g., spam vs. not spam for e-mail (cf. Jurafsky & Martin, 2009; Dickinson et al., 2013).

- *Unsupervised Learning*: ML techniques aimed at discovering some hidden structure or patterns in unlabelled data.

  Example in CL: *Document Clustering* – automatically grouping documents together based on similarity, e.g., to infer the authorship of a text (cf. Hoover, 2002; Bykh, 2011)

Each prediction made by a ML system is based on some sort of information or evidence contained in the data. The pieces of information, used by an ML algorithm to produce an output are called *Features* or *Attributes* (Kubat, 2015; Witten et al., 2011; Alpaydin, 2004). The design and extraction of such features from the data is usually subsumed under the notion of *Feature Engineering* (Scott & Matwin, 1999).

- *Feature/Attribute*: a piece of evidence, a measurement used as input for an ML algorithm. The actual feature values can be continuous or discrete.
- *Feature Engineering*: The design, definition, creation and extraction of features for ML.

The performance of an ML system can heavily dependent on the features employed for the task at hand. Some features turn out to be highly informative, other do not seem to contribute anything to the predictive power of the algorithm. State-of-the-art ML algorithms are capable of weighting the evidence provided by the particular features in an appropriate way in making their decisions. We will discuss and exemplify that issue in detail in the course of this thesis.

Connections to this thesis In this work we use ML techniques to process the relevant data, discover indicative usage patterns and classify documents with respect to the L1 of the writers, i.e., to perform NLI.

Chapter 3

Related Work

In this chapter, we discuss the related work in NLI, sketching the development of the still relatively young research area. In Section 3.1, we provide an overview of the origins, and discuss the insights and limitations of the first approaches in the area. In Section 3.2, we discuss an important milestone for the research on NLI, namely, the *First NLI Shared Task* and its contribution. Finally, in Section 3.3, we sketch the current trends and developments.

3.1 Early Work on NLI

In this section we provide an overview of the early[1] work on NLI. We present how this task has been approached in the research, show the first outcomes and point out the main issues raised in the first years.

3.1.1 NLI in the CL context

To the best of our knowledge, the earliest work that can be closely associated with NLI, was the study by Tomokiyo & Jones (2001). This contribution was concerned with classifying transcripts of English speech utterances by native or non-native English speakers. This setting is rather untypical for the NLI research, where the data usually consists of *original written L2 productions* and the task

[1]By "early" we denote the work published by 2013, i.e., before the First NLI Shared Task constituting a milestone in the given research area. The shared task is discussed in Section 3.2.

is to identify the L1 of the writers (see Section 1.1). However, conceptually and methodically that study can be certainly attributed to NLI.

NLI in the most common sense as defined in Section 1.1, started to attract attention and interest in CL with the seminal work by Koppel et al. (2005). In terms of data, the authors used a subset of the *first version* of the *International Corpus of Learner English (ICLE)* (Granger et al., 2002a). It consists of essays written by non-native English speakers of a similar age (roughly in their twenties), and at a similar level of English proficiency (higher intermediate to advanced). The study was based on texts for five L1s, namely, Bulgarian, Czech, French, Russian and Spanish, each represented by 258 essays. The authors used a Support Vector Machine (SVM) classifier employing function words, character-based n-grams, rare POS bi-grams, as well as some error types (e.g., certain spelling errors) as features. Testing was performed using 10-fold cross-validation. The best accuracy of 80.2% was obtained by a system combining all of the mentioned features. Given a chance baseline of 20%, that first outcome was already notably high.

Tsur & Rappoport (2007) replicated Koppel et al. (2005) and investigated the hypothesis that the choice of words in the second language is strongly influenced by the frequency of L1 syllables. In support of their hypothesis, the authors report that an approximation using character bi-grams alone allows classification accuracy of about 66%. Since the corpus contains learner essays on several different topics, they also investigated whether the classification with such surface features is influenced by a content bias. Using a variant of the Term Frequency - Inverted Document Frequency (TF-IDF) metric, they conclude that if a content bias exists in the corpus, it only has a minor effect. However, the hypothesis suggested and explored in this paper was later questioned by Nicolai & Kondrak (2014). The authors provided results, supporting a different hypothesis, namely, that the character bi-grams are merely mirroring the differences in the word usage rather than the phonology of L1. In particular, they showed that removing the 100 most discriminative words from the training set, resulted in a significant accuracy drop for a system based exclusively on character bi-grams. Since the phonology of L1 would influence the choice of words across the lexicon, this outcome provides some evidence against the hypothesis by Tsur & Rappoport (2007).

Estival et al. (2007) used a corpus of English e-mail data incorporating three L1s, namely English, Arabic and Spanish. The authors considered a range of different demographic and psychometric traits, including the native language, for author profiling purposes. They used a wide range of features at different levels: character-based features such as frequency of punctuation marks, lexical features such as function words as well as POS, and some features at the structural level such as paragraph breaks. Employing Information Gain as feature selection technique and a Random Forest classifier, they obtained an accuracy of 84.22%. That is a quite encouraging results. However, it is rather hard to compare it to the previous research, where mainly ICLE data was used and more different L1s were incorporated in the experiments.

Wong & Dras (2009) used the *second version* of the ICLE corpus (Granger et al., 2009), which consists of 6,085 essays (3.7 Million words) written by English learners. The corpus contains texts for 16 different L1s. This corpus became the standard data set for further NLI research. The authors compiled a subset consisting of seven L1s, namely Bulgarian, Czech, French, Russian, Spanish, Chinese and Japanese. Each L1 was represented by 70 essays for training and 25 essays for testing (plus 15 additional essays for development). On the one hand, they employed lexical features, such as function words, frequently used character-based uni-, bi-, tri-grams as well as rare and most frequently used POS bi- and tri-grams. On the other hand, they utilized three syntactic error types as features: misuse of determiners as well as subject-verb and noun-number disagreement. Employing an SVM classifier, they reported an accuracy of 73.71%. Extrapolating to a larger training set, the authors argue that this result is consistent with the findings reported by Koppel et al. (2005). However, the syntactic features utilized in their study did not improve the results obtained by using lexical features only.

Wong & Dras (2011) extended their previous work by exploring more general syntactic features based on parse trees, namely horizontal slices as well as cross-sections of such trees. These syntactic features were used in combination with lexical features employed by Wong & Dras (2009). Employing the same data set as Wong & Dras (2009) and using a Maximum Entropy classifier, the authors obtained an accuracy of up to 81.71%. The explorations show that incorporating some more sophisticated syntactic features can indeed improve the results.

Brooke & Hirst (2011) conducted several experiments using two different corpora, namely the ICLE and the Lang-8. The second corpus was compiled by the authors themselves based on the data available at `http://lang-8.com`. This data source consists of short personal journal entries of different kinds (personal narratives, requests for translations of particular phrases, etc.), which are posted by English learners in order to obtain feedback from native speakers. Compared to the ICLE corpus, there is a disproportionately high number of contributors from Eastern Asia, the level of English proficiency seems to be significantly lower, and little is known about the context of the writing for Lang-8 (e.g., there is no specification of time or resources used). To obtain texts from Lang-8 which are comparable in size to those in ICLE, Brooke & Hirst (2011) created texts consisting of multiple Lang-8 entries. The authors used character, word, and POS-based uni- and bi-grams (excluding proper nouns in case of word-based n-grams) as well as some function words as features. The experiments were based on a dataset consisting of seven L1s, namely Chinese, Japanese, French, Spanish, Italian, Polish and Russian, with each of them represented by 200 texts, drawn from each of the two corpora. The authors conducted experiments using an SVM classifier in both, a single-corpus (using 10-fold cross-validation) and a cross-corpus (training on the one corpus, testing on the other) settings. The single-corpus evaluation yielded a best accuracy of 93.8% on ICLE, and 87.7% on Lang-8 data. The results for the cross-corpus evaluation were notably worse, between 22.0% and 46.1%, whereby the best result using ICLE for training and Lang-8 for testing was at 29.3%. Based on these outcomes, the authors argue that a strong content bias is present in ICLE, which allows for an easy classification by topic rather than by L1. They thus question the general appropriateness and usefulness of that data set for the task of NLI. The authors also started investigating the usefulness of artificial learner corpora, which they compiled using machine translation of native language data. The best reported result is 67.3% in a setup with two L1s. That research direction was further explored in Brooke & Hirst (2012a).

Brooke & Hirst (2012a) extended their previous work, and showed that using automatically translated word bi-grams in combination with a new L1-transfer metric yields up to 48.3% with four L1s, when tested on ICLE. This is far below the previously reported results, but the approach promises a low content bias.

Brooke & Hirst (2012b), first of all, focused on further exploring the cross-corpus performance in NLI. The authors proposed a bias adaption technique for cross-corpus settings, which uses a small amount of data from a test corpus to overcome differences in genre or proficiency, etc. For their experiments, they compiled a data set, incorporating seven different L1s, using ICLE and Lang-8 corpora utilized before, as well as a new corpus, namely, FCE – *First Certificate in English* portion of the *Cambridge Learner Corpus* (Yannakoudakis et al., 2011). For each L1 they utilized 1000 texts from Lang-8, 200 texts from ICLE and 50 texts from FCE, plus some small amount of additional data required to apply the bias adaption technique. The authors employed a range of lexical (e.g., word n-grams) as well as syntactic (e.g., POS n-grams and CFG productions) features, utilizing SVM and Maximum Entropy classifiers. The findings supported the outcomes in Brooke & Hirst (2011), again showing that in cross-corpus settings, models trained on ICLE perform worse than models trained on Lang-8, thus questioning the use of ICLE in the context of NLI.

Tetreault et al. (2012) used four corpora for their experiments. First, the ICLE-NLI corpus, which constitutes a cleaned up version of the ICLE corpus used in the previous research. The authors removed some confounding patterns related to character encoding errors, annotations, etc., which occur predominantly in essays with certain L1s. To reduce potential topic bias issues, they discarded essays on topics, which were found for only one L1[2]. The remaining three corpora are based on the essays written by non-native English speakers in the context of the high stakes college-entrance test TOEFL®. This data set was novel. It was introduced, first of all, to overcome some apparent issues with the ICLE data, discussed in the previous research, namely the rather small data size and potential topic bias. The three TOEFL corpora differ in terms of size and/or included L1s. The main one is the *TOEFL11* corpus. It consists of 1000 essays per L1, sampled from 8 topics along with writer proficiency scores. It covers 11 L1s, namely, Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish. The other two corpora are called *TOEFL7*, which includes only the 7 L1s frequently used for NLI with ICLE; and *TOEFL-Big*, which consists of 7,900-7,983

[2]Except for Japanese, where most of the topics were unique for that L1 and removing the corresponding essays would mean to exclude Japanese from consideration.

(instead of 1000) essays per L1 for the same 11 L1s as in TOEFL11, with the data sampled from a wider range of topics. The authors utilized features, such as character-based n-grams, POS-bi-grams, word-based uni- and bi-grams, spelling errors, some syntactic features based on TSGs and Stanford Dependencies as well as language models. They employed Logistic Regression as classifier. The best accuracy on the ICLE-NLI corpus was 90.1%, applying 5-fold cross-validation. The ICLE-trained models did not generalize well to the TOEFL11 data, which is in line with the cross-corpus findings presented by Brooke & Hirst (2011, 2012b). Nevertheless, the authors conclude that many of the trends found in the previous work using ICLE, do well generalize to other corpora, e.g., the power of the features introduced by Koppel et al. (2005), and the high performance of word-based n-grams as features, which showed reasonable accuracies even for a corpus, specifically designed to overcome topic bias issues. The authors also show that corpus size have an effect on the classification performance: Models trained on small corpora do not seem to generalize well for the task at hand. They also explored combining the different features employing a probability-based ensemble classifier instead of using a simple flat model, where all features are put in a single vector. The ensemble classifiers consistently outperformed the simple models. Finally, the authors also showed that language proficiency can have some effect on the NLI performance, stating that further research is required to better understand that phenomenon.

Krivanek (2012) used the ICLE corpus and the LOCNESS[3], a corpus consisting of 366 essays written by native English speakers. In this study, the author employs some linguistically motivated features encoding language variation, namely, theory- and data-driven verb alternations based on CFG parses. The approach showed first promising results. In particular, the system yielded an accuracy of up to 83.3% in a binary classification task using L1 Chinese and native English texts, with 300 essays for training and 60 essays for testing per L1. Moreover, the work provides some first interesting qualitative insights regarding the preferences for particular syntactic options by English learners with different L1s. For example, it turned out that in the context of the "Locative preposition drop alternation" (e.g., *Martha climbed up the mountain* vs. *Martha climbed the mountain*), L1 Chi-

[3] `https://www.uclouvain.be/en-cecl-locness.html`

nese learners of English tend to prefer the variant without the locative preposition compared to the native English speakers.[4]

Bykh & Meurers (2012) is our initial contribution to NLI. The study was based on a ICLE data set incorporating seven different L1s following the Wong & Dras (2009) setup. It was concerned with a broad systematic exploration of recurring word-, POS- and OCPOS-based n-grams up to the maximum length as features, as well as a contribution to the cross-corpus performance discussion, started by Brooke & Hirst (2011). Different from related research discussed above, our findings suggest that ICLE-trained models can still yield a reasonable cross-corpus performance. We present the details and the results of the approach in Chapter 5.

3.1.2 NLI in the SLA context

Besides the explorations on NLI in CL, there have been some related contributions in the research on SLA. A particular comprehensive and representative work is Jarvis & Crossley (2012). It combines five studies on NLI with the focus on exploring different feature types, and interpreting the outcomes from the qualitative point of view in the context of the SLA research. Most of these contributions are based on the ICLE data. All of them employ Linear Discriminant Analysis (LDA) as classification method and use cross-validation. In the following we briefly sketch these studies.

Jarvis et al. (2012b) used uni-grams to investigate the degree to which learners' word-choice patterns reflect L1-specific lexical use tendencies. Different from the other studies presented in Jarvis & Crossley (2012), which make use of the ICLE corpus, here the authors utilized a corpus of written narrative descriptions of a silent film, namely, the eight-minute "Alone and Hungry" segment of Chaplin's "Modern Times", compiled by the authors themselves. That corpus consists of 446 texts, produced by learners of English with five different L1s, namely, Danish, Finnish, Portuguese, Spanish and Swedish. The classification accuracy was at 76.9% using 53 features (most frequent words across the L1s). The authors also provide some interesting qualitative analysis, e.g., they found out that speakers

[4]That contribution is of high relevance in the context of this thesis and is discussed in more detail in Section 7.3.

with L1 Finnish show the lowest use of *he* and *she*. They explain it by the fact that Finnish is the only L1 used in the study, which lacks pronomial gender, and thus, as consequence these speakers seem to resort to the use of referential nouns more often in their L1 than others, which also seems to transfer to the L2 production. Another example is the underuse of the articles by writers with L1 Finnish, which can be attributed to the fact that Finnish lacks articles as a grammatical class.

Jarvis & Paquot (2012) extended the approach by Jarvis et al. (2012b) and used a bigger set of some most frequent n-grams with $1 \leq n \leq 4$ (the topic and prompt-specific words were omitted), based on ICLE data, incorporating 12 different L1s. The best result of 53.6% was obtained by using all of the different n-gram lengths in combination with feature selection (first of all, in order to get feature numbers suitable for the LDA classifier). The best performing single n-gram length was $n = 1$ (uni-grams) yielding an accuracy of 53%. Regarding the qualitative analysis, one of the interesting examples provided by the authors is the bi-gram *going to*, which seems to be highly indicative for the learners with L1 Spanish. The authors suggest that it probably originates from the frequent usage of the pattern *ir a + INFINITIVE* in L1 Spanish, indicating future intentions. Another example is the overuse of the tri-gram *all the time* by the writers with L1 Finnish, which can be presumably linked to the common Finnish phrase *koko ajan* (lit. *whole time*), covering a range of meanings from *usually* to *most/all of the time*.

Crossley & McNamara (2012) used more abstract features, such as coherence and lexical richness, obtained using the Coh-Metrix tool (Graesser et al., 2012). For their experiments they employed the ICLE data, utilizing essays with four different L1s, namely, Czech, German, Finnish and Spanish. The best accuracy was 66% using 14 features. The authors provide an interesting characterization of the writers with different L1s based on the explored features. For example, they state that writers with L1 German seem to produce texts with lower lexical cohesion, and they use in general more concrete, more meaningful, more familiar, less specific and less ambiguous words. Whereas the writes with L1 Spanish show rather high degrees of lexical cohesion, and they tend to produce frequent words, which are rather lexically sophisticated, i.e., words with low concreteness and familiarity scores, etc.

Bestgen et al. (2012) used features based on error categories, namely, formal errors, grammatical errors, lexical errors, lexico-grammatical errors, style errors, punctuation errors, and errors based on word occurrence (redundant, missing or misordered words). In sum, they employed 46 error subcategories based on manually error-tagged ICLE subset. The utilized data consisted of 223 essays including three different L1s, namely, French, German and Spanish. It turned out that only 12 out of the 46 features showed significant differences across the L1s. The best classification accuracy was at 65%. The authors argue that not all of the errors necessarily resulted from transfer effects, but some of them can indeed be attributed to L1-transfer. For example, writes with L1 Spanish tend to omit personal pronouns in the subject position, which is a valid pattern in Spanish, but not in French or German. Similarly, the usage of *in* instead of *on* in Spanish L1 writings can be attributed to the fact that both prepositions are translated into Spanish by the same preposition *en*.

Jarvis et al. (2012a) combines the features used in all the different contributions included in Jarvis & Crossley (2012), presented above – features based on n-grams, Coh-Metrix and error tags. For that comparative study the authors employed the data used in Bestgen et al. (2012). First, they tested the performance of the separate models on the given data, and reported comparable accuracies for the different models, namely, around 63–65%. The authors conclude that none of them perform significantly better than the others. However, combining the models and performing feature selection yielded an accuracy of 79.4%, which significantly outperformed each of the separate models. Thus, the authors conclude that combining different types of evidence is beneficial for the task at hand, which is well in line with the previous research on NLI.

3.1.3 Summary

In this section we provided an overview of the early research on NLI. We discussed several contributions in detail, including both, studies in CL and SLA research. The CL contributions were mainly focused on tuning the models and increasing the performance of the ML systems, whereas the contributions from the SLA research showed increased interest in the qualitative analysis of the re-

sults in the context of L1-transfer. Hence, both research strands completed each other and contributed many interesting and valuable insights, and thus, brought the research field of NLI a good way forward.

NLI was usually approached as a text classification problem with the different L1s as class labels. The most widely used features were n-grams of different types and various lengths, but also more linguistically motivated features, such as different error types and features based on CFG rules, TSG's and dependencies. The most preferred classifier was a Support Vector Machine (SVM).

The main issues and concerns in the research on NLI were related to the used data. Because of a scarcity of appropriate corpora, almost all of the studies were based on the *International Corpus of Learner English* (ICLE). However, that corpus consists of only roughly six thousand texts distributed across 16 different L1s, which seems rather small for training and testing statistical systems. Hence in general it is hard to assess the scalability and generalizability of the findings. Moreover, it was suggested that this data seems to suffer from issues such as topic bias, yet another factor potentially compromising the general validity of the outcomes. Because of these issues with the ICLE data, some researchers started using some new corpora, such as the Lang-8, FCE or TOEFL11 for their evaluations. That was an important step for advancing the generalizability of the findings. Another big issue is the general comparability of the findings provided by the various contributions at that early stage. Even if the studies were based on the same corpus, they are still difficult to compare because of differences in some parameters. For example, the contributions used:

- different sizes of the data subsets
- different numbers of L1s
- different L1s
- potentially different texts for the same L1s
- different preprocessing, etc.

These factors inevitably lead to different headaches, whenever one tried to compare the findings. Fortunately, there was organized an event, aimed to cope

with most of the main issues in the field, and that eventually became an important milestone for NLI: *The First NLI Shared Task* / *NLI Shared Task 2013*, which we discuss in the next section.

3.2 The Contribution of the First NLI Shared Task

The *First Native Language Identification Shared Task* (First NLI Shared Task / NLI Shared Task 2013) was an event hosted at *The 8th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-8)*, associated with *The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*. The detailed report on the task, its context and the results are provided in Tetreault et al. (2013), which is the contribution we base our description and our discussions in this section on.

The aim of the First NLI Shared Task (*FNLIST*) was to cope with the core issues in the research on NLI, presented and discussed in the Section 3.1, e.g., the small data sizes the systems were trained on, or some idiosyncrasies such as topic bias, apparently skewing the results.

First, since most of the issues were directly related to the data used for NLI in the previous research, the solution must have included an improvement on this point. Thus, a highly important contribution of the FNLIST was the release of a new data set to the community, namely, the TOEFL11 corpus (Blanchard et al., 2013). This corpus was designed specifically for the task of NLI. Essentially, it constitutes an extended version of the data, introduced and used by Tetreault et al. (2012). Second, the shared task provided evaluation standards, which all participants had to comply with, thus enabling direct and detailed comparisons of different approaches. This was nearly impossible in NLI before (see Section 3.1), but clearly crucial for advancing the research in any area.

In the following sections, we describe the version of the TOEFL11 corpus released for the FNLIST, the different tasks and the obtained results. Finally, we summarize the contribution of the FNLIST.

3.2.1 Data

The *TOEFL11* corpus alias *The ETS Corpus of Non-Native Written English* (Blanchard et al., 2013) consists of essays written by English learners in the context of the high-stakes college-entrance test TOEFL® (*The Test of English as a Foreign Language*). It includes data on 11 different L1s, namely, Arabic (ARA), Chinese (CHI), French (FRE), German (GER), Hindi (HIN), Italian (ITA), Japanese (JPN), Korean (KOR), Spanish (SPA), Telugu (TEL) and Turkish (TUR). Each of the 11 L1s is represented by 1,100 essays as follows – training set: 900 essays, development set: 100 essays, test set: 100 essays. Thus, in sum there are three subsets, namely, TOEFL11 *train* (9,000 essays), TOEFL11 *dev* (1,100 essays) and TOEFL11 *test* (1,100 essays) amounting to overall 12,100 essays. The texts were sampled as evenly as possible from eight prompts. The data is categorized by the proficiency levels, namely, *low*, *medium* and *high*.

3.2.2 Tasks

The FNLIST consisted of three tasks. The *test data* was always the same, namely, the TOEFL11 *test* set. The *training data* varied across the tasks. The task identifiers and the training data, permitted to use for the individual tasks is listed below:

1. *Closed* task: TOEFL11 *train*, TOEFL11 *dev*
2. *Open-1* task: any data, excluding: TOEFL11 *train*, TOEFL11 *dev*
3. *Open-2* task: any data

The *closed* task was the main task with 29 participating teams from around the globe. Since the same data was used for training and testing the systems, all results are directly comparable. The number of participants for the *open-1* task was three, and for the *open-2* – four. The results for the two *open* tasks are not directly comparable to the *closed* task or to each other, because of the differing training sets used by the different teams. Nevertheless, the corresponding findings are interesting and highly relevant in the context of the discussion on cross-corpus evaluation in NLI.

3.2.3 Approaches and Results

In this section we present a brief summary of the common approaches as well as the results of the FNLIST based on the official report (Tetreault et al., 2013).

Classifiers The most preferred classifier was SVM, followed by Logistic Regression / MaxEnt, which is well in line with the previous research. Some teams experimented with ensemble methods. Others employed Discriminant Function Analysis or k-NN, etc.

Features In this paragraph we list the most commonly used feature types.

1. N-grams of various types and lengths, namely:
   - Character-based, $1 \leq n \leq 9$
   - Word-based, $1 \leq n \leq 5$
   - POS-based, $1 \leq n \leq 5$
2. Syntactic features, based on:
   - Dependencies
   - CFG production rules
   - Tree Substitution Grammars
3. Spelling features

Other features used by some teams included function n-grams, complexity measures, suffix-based features, and features based on Adaptor Grammars, etc.

Results The overall results for the three tasks, listing all teams (along with the corresponding team identifiers) participating in each task, are presented in the Tables 3.1, 3.2 and 3.3. The chance baseline for all tasks is 9.1%.

Among the participants was the team *Tuebingen (TUE)*, supervised by Prof. Dr. Detmar Meurers an lead by the author of this thesis. The accuracies obtained by our systems are highlighted in boldface in each of the tables. We discuss

rank	team	accuracy
1	Jarvis (JAR)	83.6%
2	Oslo NLI (OSL)	83.4%
3	Unibuc (BUC)	82.7%
4	MITRE "Carnie" (CAR)	82.6%
5	**Tuebingen (TUE)**	**82.2%**
6	NRC (NRC)	81.8%
7	CMU-Haifa (HAI)	81.5%
8	Cologne-Nijmegen (CN)	81.4%
9	NAIST (NAI)	81.1%
10	UTD (UTD)	80.9%
11	Ualberta (UAB)	80.3%
12	Toronto (TOR)	80.2%
13	MQ (MQ)	80.1%
14	cywy (CYW)	79.7%
15	dartmouth (DAR)	78.1%
16	ItaliaNLP (ITA)	77.9%
17	Chonger (CHO)	77.5%
18	HAUTCS (HAU)	77.3%
19	LIMSI (LIM)	75.6%
20	CoRAL Lab @ UAB (COR)	74.8%
21	LTRC IIIT Hyderabad (HYD)	74.4%
22	CUNI / Charles University (CUN)	72.5%
23	UNT (UNT)	64.5%
24	Bobicev (BOB)	62.5%
25	kyle, crossley, dai, mcnamara (KYL)	59.0%
26	UKP (UKP)	58.3%
27	Michigan (MIC)	43.0%
28	eurac (EUR)	38.6%
29	VTEX (VTX)	31.9%

Table 3.1: Results of the FNLIST *closed* (main) task.

our contribution, namely, Bykh et al. (2013), in detail in Section 6. In sum, our systems scored as follows:

- *Closed* task: rank five / 29 participating teams
- *Open-1* task: rank two / three participating teams
- *Open-2* task: rank one, i.e., winner / four participating teams

***Closed* (main) task** The best accuracy in the main task was 83.6%, achieved by the team *JAR* (Jarvis et al., 2013, see Table 3.1). That result was obtained using word- (lexeme-), lemma- and POS-based n-grams with $1 \leq n \leq 3$, which occurred in at least two texts of the training set[5], and applying log-entropy weighting schema as well as normalizing each text to the unit length. As classifier the authors employed a L1-regularized L2-loss SVM using the LIBLINEAR package (Fan et al., 2008). In sum, having an evaluation context, which finally ensures direct comparability of a range of approaches, and considering many of the top scoring submissions, we can state that lexical traits captured by various n-gram types, constitute the best performing features for NLI (e.g., Jarvis et al., 2013; Lynum, 2013; Henderson et al., 2013; Bykh et al., 2013; Gebre et al., 2013, etc.).

***Open-1* task** This is the genuine cross-corpus task in the context of FNLIST, where any data excluding TOEFL11 was permitted to use for training the systems. Thus, a critical point here was getting hold of sufficient non-TOEFL11 training data. The employed feature sets were based on what was used in the *closed* task.

rank	team	accuracy
1	Toronto (TOR)	56.5%
2	**Tuebingen (TUE)**	**38.5%**
3	NAIST (NAI)	35.6%

Table 3.2: Results of the FNLIST *open-1* task.

The best accuracy was 56.5%, achieved by the team *TOR* (Brooke & Hirst, 2013, see Table 3.2). The authors used word-based and mixed POS/Function n-grams as well as features based on dependencies. The best submitted system was based on employing the following corpora: ICLE (Granger et al., 2009), FCE (Yannakoudakis et al., 2011), Lang-8 (Brooke & Hirst, 2011, 2012b), and also some new corpora, i.e., the *International Corpus Network of Asian Learners of English (ICNALE)* (Ishikawa, 2011), as well as some Indian news articles from Hindi and Telugu newspapers written in English. The authors also explored using translated versions (by Google Translate) of Indian blogs from the *ICWSM 2009 Spinn3r Dataset* (Burton et al., 2009) and some Tweets from the *WORLD twitter*

[5]I.e., *recurring* n-grams, cf. Bykh & Meurers (2012).

corpus (Han et al., 2012), geolocated in Hindi and Telugu speaking areas. In further post-hoc explorations, some systems including the Twitter data even slightly outperformed their best submission. The authors also state that using the bias adaption technique presented in Brooke & Hirst (2012b) consistently provided an quantitative advantage. For our submission (*TUE*), which scored second, we used besides the ICLE, FCE and ICNALE corpora mentioned above, also the *BUiD (British University in Dubai) Arab Learner Corpus (BALC)* (Randall & Groom, 2009), as well as an own corpus for Telugu, we call *Tübingen Telugu NLI Corpus (TÜTEL-NLI)*, consisting of English texts written by Telugu native speakers from bilingual (English-Telugu) blogs, literary articles, news and movie review websites. Finally, the team *NAI* employed only the Lang-8 data in the given context. The findings show that using more training data, even if it is out of domain, usually yields higher accuracies. The results suggest that high cross-corpus performance still remains challenging in NLI.

***Open-2* task** In this task, where the usage of any data was permitted for training the systems, the training sets utilized for the *closed* and *open-1* tasks were combined to train the models. Again, the feature sets were based on what was used in the *closed* task.

rank	team	accuracy
1	**Tuebingen (TUE)**	**83.5%**
2	Toronto (TOR)	81.6%
3	LTRC IIIT Hyderabad (HYD)	74.1%
4	NAIST (NAI)	70.3%

Table 3.3: Results of the FNLIST *open-2* task.

The best accuracy of 83.5%, was achieved by our team *TUE* (Bykh et al., 2013, see Table 3.3). We used for all tasks the same feature set, consisting of a wide range of surface-based and linguistically-motivated features, such as word- and POS-based recurring n-grams, linguistic complexity features, features based on dependencies, CFG production rules (local trees) and suffixes, etc. The different feature types were combined using a probability-based ensemble classifier. A detailed description and discussion of our feature set and the employed classifier is

provided in Section 6. The findings suggest that adding some additional data to the TOEFL11 training set, has the potential to further improve the already relatively high accuracies from the *closed* task.

3.2.4 Summary

To sum up, the *First NLI Shared Task* was a milestone for the research on NLI. It brought together 29 teams from around the globe and significantly advanced the research in the given area. It provided a new data set to the community, which was specifically designed for the task of NLI, overcoming most of the critical issues with the previously used corpora. The shared task facilitated a meaningful comparison of many different approaches to NLI employing various features and classification techniques. Moreover, it facilitated the comparability of future contributions by providing a standard evaluation setup. The results showed that high accuracies, up to 83.6%, are possible for distinguishing between 11 L1s. At the same time, the findings suggest that obtaining similar cross-corpus results is still challenging, confirming the previous research. Furthermore, the outcomes show that lexical features such as n-grams seem to be most powerful. These were consistently among the best performing features in both, single- and cross-corpus settings. After the shared task, the state-of-the-art in NLI became more clear, which smoothed the way for the future research. In the following section we sketch the current trends in NLI.

3.3 Current Trends in NLI

After the discussion of the early research on NLI and the contribution of the First NLI Shared Task as an important milestone in the given area, in this section we turn to sketching some current research developments. In sum, we can observe the following trends:

1. Exploring further cross-corpus settings (section 3.3.1);

2. Exploring L2s different from English (section 3.3.2);

3. Exploring new features and techniques (section 3.3.3);

4. Advancing qualitative analysis (section 3.3.4);

3.3.1 Cross-corpus Evaluation

The previous research on NLI showed that achieving a cross-corpus performance, comparable to the relatively high single-corpus results obtained so far, remains a challenging task. Since it is highly desirable to have robust, domain independent NLI systems fostering generalizable findings, many recent contributions included cross-corpus experiments.

Some researchers extended the data sets by new corpora, e.g., Malmasi & Dras (2015a) employed the *EF Cambridge Open Language Database (EFCamDat)* corpus (Geertzen et al., 2013, 2014), previously used only for single-corpus NLI experiments (Jiang et al., 2014). Other authors conducted new experiments and evaluations employing corpora such as TOEFL11, ICLE and FCE, etc., used before for a range of single-corpus as well as for some cross-corpus experiments (see Sections 3.1 and 3.2).

The cross-corpus evaluations were performed based on features previously used for NLI, such as n-grams, CFG production rules and TSGs, or modified versions of those features, which were usually employed in connection with some new classification and evaluation techniques[6] (Malmasi & Dras, 2015b; Bykh & Meurers, 2014; Ionescu et al., 2014; Swanson & Charniak, 2013, 2014).

In sum, the experiments show that the cross-corpus results remain far below single-corpus, suggesting that high cross-corpus performance still remains one of the important tasks, to be targeted in the future work on NLI.

3.3.2 L2s Different from English

English has become the dominant global language of communication over the second half of the 20th century. There are about 400 Million native speakers in Britain, the United States and the Commonwealth, as well as over a Billion

[6]We briefly discuss the new techniques in the Sections 3.3.3 and 3.3.4.

non-native speakers around the world (Guo & Beckett, 2007). Thus, it is not surprising, that until recently the work in many NLP areas in general, and on NLI in particular, was essentially focused solely on English as L2. Golcher & Reznicek (2011) was an exception, employing L2 German in the context of NLI. However, recently there was a shift towards more diversity in that regard, with several contributions exploring L2s different from English. Among the employed L2s were, e.g., Arabic, Chinese, Czech, Finnish and Norwegian (Aharodnik et al., 2013; Malmasi & Dras, 2014b,a,c; Malmasi et al., 2015a). The results are promising, showing accuracies well above the chance baseline, i.e., up to around 70–80% for 7–11 L1s, which is comparable to the performance for L2 English (see Section 3.1 and Section 3.2). However, working with languages different from English poses some additional issues, such as the lack of reasonably sized data sets for training robust statistical models, and the shortage of NLP tools for identifying different linguistic features. For example, for the Finnish NLI experiments reported by Malmasi & Dras (2014c), the authors were only able to use 12–40 texts per L1, and Malmasi et al. (2015a) state the rather limited availability of NLP tools for Norwegian, which hinders further investigations, etc.

Exploring L2s different from English remains an important and interesting research direction, enabling a more broad and general view on NLI.

3.3.3 New Features and Techniques

A wide range of different features and techniques have been explored for NLI in the context of the First NLI Shared Task (see Section 3.2). Nevertheless, some recent contributions show that there still remains sufficient space for further explorations. Some of the recently reported systems outperformed the best result of the First NLI Shared Task following its standard setup (Bykh & Meurers, 2016, 2014; Ionescu et al., 2014)

In Bykh & Meurers (2014) we further explored CFG production rules as features. We included phrasal as well as lexicalized categories, and employed feature encodings, which are inspired by the variationist perspective (see Section 2.2). Furthermore, we explored combining different models utilizing an probability-based ensemble classifier, which uses the estimates yielded by different individual

classifiers as features (Tetreault et al., 2012). Moreover, we investigate applying some ensemble optimization and tuning techniques. The proposed linguistic features alone yielded accuracies up to 79.6%, and a combination with four types of n-grams showed a best result of 84.8%, thus outperforming the winning systems of the First NLI Shared Task by 1.2%. We will discuss this contribution in more detail in Chapter 9 and Part IV.

Ionescu et al. (2014) employed character n-grams as features, and explored using several string kernels and their combinations via multiple kernel learning. They report a best accuracy of 85.3%, which outperforms the winning system of the First NLI Shared Task by 1.7%.

In Meurers, Krivanek & Bykh (2014) we discuss the importance of linguistic generalizations for tasks such as NLI. First, we used modified versions of word-based n-grams, and showed that linguistic abstraction using POS, when applied to parts of such n-grams, can increase the performance of the system. Whereas, applying some non-linguistic abstraction using simple wildcards, decreases the accuracy of the system. Thus, it shows that using POS information can be superior to using words. Second, based on Krivanek (2012) we present some results on employing theory- and data-driven verb alternations based on CFG parses as features. The approach showed first promising results for a binary classification task aimed at distinguishing L1 Chinese learners from native English speakers (see Section 7.3 for more details).

Malmasi et al. (2015b) explored different ensemble methods, including an oracle, to estimate the upper-bound of accuracy for the task of NLI. Combining all of the submissions to the First NLI Shared Task, the oracle shows an impressive accuracy of 99.5%. This poses a result yielded by a (hypothetical) optimal system, always making the correct prediction, given that such a prediction is provided by any ensemble member. While it is interesting to see what is potentially possible, the authors conclude that the actual ceiling could be substantially lower. However, the technique can also be used to isolate the subset of texts for further analyses, which are consistently hard to classify correctly.

In Bykh & Meurers (2016) following Meurers, Krivanek & Bykh (2014), we explored dependency-based verb subcategorization features under a variationist perspective, considering the given verb lemmas as linguistic variables and the

subcategorization options as variants. We propose a method based on hierarchical clustering, capable of abstracting from individual variables to classes. On the one hand, we showed that using the grouping technique can reduce data sparsity and make the models more compact and efficient. Combining the new verb subcategorization features with some features used before (Bykh & Meurers, 2012; Bykh et al., 2013; Bykh & Meurers, 2014), we obtained an accuracy of 85.4% – To the best of our knowledge, this was the best result published by then for the standard TOEFL11 data setup. On the other hand, we showed that the method can advance the qualitative analysis in NLI by facilitating the validation of hypothesis about L1-transfer. These findings contribute to the strand of research pursued in some recent publications (Swanson & Charniak, 2013, 2014; Malmasi & Dras, 2014d), suggesting techniques for advancing the qualitative analysis in NLI. This poses another interesting trend in the research, which we discuss in the next section.

In sum, we can observe a range of new interesting approaches after the First NLI Shared Task, showing that the research area still remains vivid.

3.3.4 Qualitative Analysis

Most of the contributions on NLI primarily focus on advancing the quantitative aspect of the task, i.e., on improving the performance of the various NLI systems. However, another interesting aspect is the qualitative analysis of the findings and the development of techniques capable of advancing that research direction. It can help discovering some new L1-transfer effects, and enhance our understanding of L1-transfer in general. Thus, recently, besides the qualitative findings provided by the contributions in the SLA research (see Section 3.1), we can also observe a growing interest in advancing qualitative analysis in the CL publications on NLI (Swanson & Charniak, 2013, 2014; Meurers et al., 2014; Malmasi & Dras, 2014d, 2015a; Malmasi & Cahill, 2015).

Swanson & Charniak (2013) propose a technique for identifying linguistic patterns that learners with certain L1s use with markedly unusual frequencies. The study is concerned with syntactic patterns based on TSGs. The authors use different corpora for L2 English with four different L1s, namely, Chinese, German, Spanish and Japanese. They explore several feature relevancy measures such as

InfoGain and χ^2, as well as feature redundancy measures such as *Symmetric Uncertainty* and *Normalized Pointwise Mutual Information*. The end result of the evaluation is a list of patterns along with their usage statistics, which can be a useful resource for further SLA research. For example, the authors figured out that Spanish L1 speakers prefer using the phrase *The X of Y*, the verb *go* and the determiner *this*, the Japanese L1 speakers are seen to frequently use a personal pronoun for subject, and the German L1 speakers apparently tend to begin sentences with adverbs, etc. Swanson & Charniak (2014) adapts the approach proposed in Swanson & Charniak (2013) to dependencies, again resulting in a list of potentially interesting L1-transfer candidates.

Malmasi & Dras (2014d) proposed using the positive/negative weights assigned by an SVM classifier (LIBLINEAR, see Fan et al., 2008) to detect overused/underused patterns. The authors evaluate the approach on the TOEFL11 corpus, using features based on Adaptor grammar collocations and Stanford dependencies, as well as some lexical features. The results suggest that, e.g., L1 Chinese learners tend to underuse determiners, which is an issue well-known in the SLA research (Robertson, 2000); whereas L1 German learners tend to overuse the "existential there" construction (*there is/are*), which seems to be due to the frequent use of the equivalent *es gibt* in German, etc. Malmasi & Dras (2015a) applied the method proposed in Malmasi & Dras (2014d) in a cross-corpus setting, employing the EFCamDat and TOEFL11 corpora, and using function words, POS n-grams as well as CFG production rules as features.

In Meurers, Krivanek & Bykh (2014) based on Krivanek (2012), we discuss the usage patterns for different verb alternations employing ICLE and LOCNESS data. We show, e.g., that in the context of the "Locative preposition drop alternation" (e.g., *Martha climbed up the mountain* vs. *Martha climbed* $\emptyset$ *the mountain*), L1 Chinese learners tend to overuse the variant without the locative preposition compared to the native English speakers. Whereas, for the "Dative alternation" (e.g., *He gave John the book* vs. *The gave the book to John*) the distribution between the variants is very similar for L1 Chinese and native English speakers.

Malmasi & Cahill (2015) investigated the correlation between different features commonly used for NLI. Their findings confirm that, as assumed, syntactic and lexical features correlate least, thus apparently capturing different effects.

In Bykh & Meurers (2016) we employed the variationist perspective to features based on verb subcategorization. We proposed a method based on hierarchical clustering, capable of abstracting from individual variables to classes, and showed that it can advance the qualitative analysis in NLI by facilitating the validation of hypothesis about L1-transfer.

The qualitative analysis of the results, as well as the development of techniques fostering that process is, from our point of view, one of the most important research directions in NLI. On the one hand, such analysis can help detecting most indicative features and thus improve the performance of the NLI systems. On the other hand, it is well-capable of providing valuable linguistic insights, which can be used in the context of the foreign language teaching or language tutoring systems, etc.

3.3.5 Summary

After the First NLI Shared Task, which served as a test bed for a wide range of different approaches and explorations, the research on NLI remained vivid, with many contributions showing that it is far from becoming exhausted. Some recent approaches further focused on the quantitative side and managed to outperform the best result of the shared task. Other contributions were concerned with targeting issues, such as cross-corpus evaluation, or with further advancing the qualitative insights in SLA based on NLI. Others started exploring NLI for L2s different from English. In sum, we observe a high diversity in the research on NLI, which promises further significant advances in the field.

3.4 Summary

The task of NLI started to increasingly attract interest after the seminal work published by Koppel et al. (2005). It approached NLI as a text classification problem with the different L1s as classes, using some surface based and linguistically motivated features. In general, the subsequent contributions in CL followed the idea of that approach and provided some valuable modifications and extensions to it. That, first of all, lead to significant improvements in the classification accuracy.

However, the research also revealed some first issues, mainly related to the data used for NLI. Besides the quantitative advances, the question of the qualitative insights was brought forward by several SLA researchers contributing to the field (Jarvis & Crossley, 2012).

In 2013, the First NLI Shared Task was organized to cope with the central issues in the research on NLI. A new corpus, namely TOEFL11, specifically designed for NLI, was released to the community. Moreover, a setup which allows for a better comparison of the various NLI systems, was established in the field. In sum, 29 teams from around the globe participated in the shared task, further advancing the research on NLI.

The research area remains vivid, showing some interesting trends, such as advancing the generalizability of the findings by the cross-corpus evaluation, investigating NLI for L2s different from English, exploring novel features and techniques, as well as extending the qualitative analysis based on the NLI outcomes.

Part II

Broad Linguistic Feature Exploration

Chapter 4

Introduction

In this part of the thesis, we focus on the quantitative exploration of a broad range of surface-based and linguistically-motivated features for the task of NLI. Our aim is to investigate their use for NLI in terms of classification accuracy. We start by systematically exploring a range of features based on different types of n-grams in Chapter 5, and continue by broadening the perspective utilizing a range of more linguistically-motivated features in Chapter 6.

Chapter 5

Systematically Exploring Recurring N-grams as Features

5.1 Introduction

Inspired by the variation n-gram approach to corpus annotation error detection (Dickinson & Meurers, 2003, 2005; Boyd et al., 2008), and the successful application of n-grams re-occurring in the data for authorship attribution (Bykh, 2011), we discuss the use of recurring n-grams of any length as features for training an NLI system. Most of this chapter reports our results published in Bykh & Meurers (2012). In addition, it also provides some of our findings published in Meurers et al. (2014).

The rest of this chapter is organized as follows: First, in Section 5.2, we systematically explore the performance of word-based n-grams as well as n-grams incorporating two degrees of abstraction using POS information. Second, in Section 5.3, we investigate the generalizability of our results across corpora. In particular, we investigate the claim by Brooke & Hirst (2011) that due to topic bias, models trained on ICLE do not seem to generalize to other corpora. We show that training our model on ICLE and testing it on three other, independently compiled learner corpora dealing with other topics, still results in a reasonably high classification accuracy. Finally, in Section 5.4, we explore the effect of applying a linguistic and non-linguistic generalization to word-based n-grams as features.

5.2 Systematic Feature Exploration

5.2.1 Data

For the first, core study in this chapter, we use a subset of the *International Corpus of Learner English* (ICLE, second version; Granger et al., 2009). The overall ICLE corpus consists of 6,085 essays written by English learners with 16 different L1s. They are at a similar level of English proficiency, namely higher intermediate to advanced and of about the same age. Following the setup of Wong & Dras (2009), we randomly select a set of essays from the same seven L1s – namely, Bulgarian, Czech, French, Russian, Spanish, Chinese, and Japanese – and we use the same data split with 70 essays for training and 25 essays for testing for each of the L1s. This results in a total of 490 essays for training and 175 for testing. As in Wong & Dras (2009), we only included essays between 500 and 1000 words in length. We tokenized the data and removed all punctuation marks, special characters and capitalization. Thus, each essay is represented as an array of lower-case words.

To get a better sense of how well our approach performs, we conducted ten experiments. We select the data for each of them randomly from the full set of ICLE essays within the mentioned length range. We thus are able to observe the variance of the results based on ten randomly selected samples from the overall corpus subset matching the described criteria. We first describe one of the ten experiments in detail and then turn to the overall ten experiments.

5.2.2 Features

Different from previous research, in this study we explore *recurring n-grams of all occurring lengths* as classifier features. By *recurring* we here mean all n-grams that occur in at least *two* different essays of the training set d.[1] *Of all occurring lengths* means all recurring n-grams up to the maximum possible n value occurring in d, i.e., all n-grams with $1 \leq n \leq max_n(d)$.

[1]Later contributions, in particular Jarvis et al. (2013), which was the winning system of the First NLI Shared Task, used various types of recurring n-grams for successfully maximizing the classification performance in NLI (see Section 3.2).

On the one hand, we use recurring *word-based* n-grams directly, i.e., the surface forms. On the other hand, we explore two different classes of recurring *POS-based* n-grams as a generalization, based on a POS tagged version of the corpus using the *PennTreebank* tagset (Santorini, 1990)[2]. In sum, we define our features based on the following three classes of recurring n-grams:

Word-based n-grams (word n-grams): strings of words, i.e., the surface forms

- $n = 1$: *analyzing, attended, ...*
- $n = 2$: *aspect of, could only, ...*
- $n = 3$: *is capable of, the assumption that, ...*
- ...

POS-based n-grams (POS n-grams): *all words* are converted to the corresponding POS tags

- $n = 1$: *nnp, md, nns, vbd, ...*
- $n = 2$: *nns md, nn rbs, nn rbr, cc wdt, vbp jjr, vbp jjs, ...*
- $n = 3$: *cd wdt md, vbp nn md, dt rbr cc, nn jj in, ...*
- ...

Open-Class-POS-based n-grams (OCPOS n-grams)[3]: *nouns, verbs, adjectives* and *cardinal numbers* are converted to the corresponding POS tags

- $n = 1$: *far, vbz, much, jj, ...*
- $n = 2$: *nn whenever, jj well, jjs vbd, vbg each, nn always, ...*
- $n = 3$: *vbp currently jj, only to the, cd vbz jj, vb if there, ...*
- ...

We explore the whole range of n values as well as all possible $[1, n]$ intervals. Figure 5.1 depicts the counts of *different n-grams* for each n (for uni-grams, bi-grams, tri-grams, etc.) and Figure 5.2 for each $[1, n]$ interval (for uni-grams alone, uni-grams and bi-grams together, uni-grams, bi-grams and tri-grams together, etc.).

[2] `https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`

[3] Similar representations are also used by Baroni & Bernardini (2006) for the identification of "translationese".

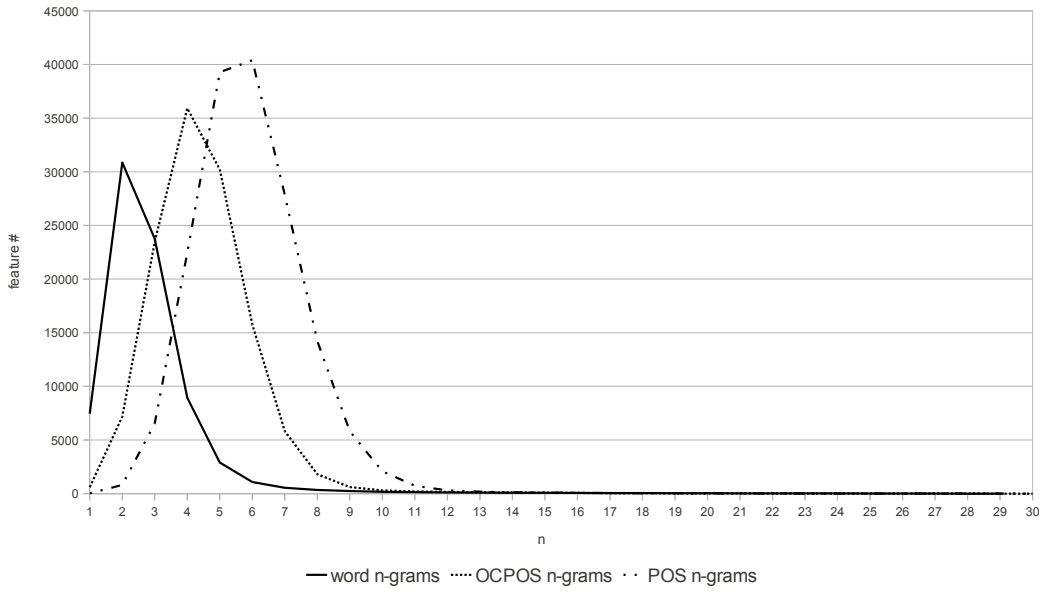

Figure 5.1: Feature counts for single n n-gram settings for the single ICLE sample

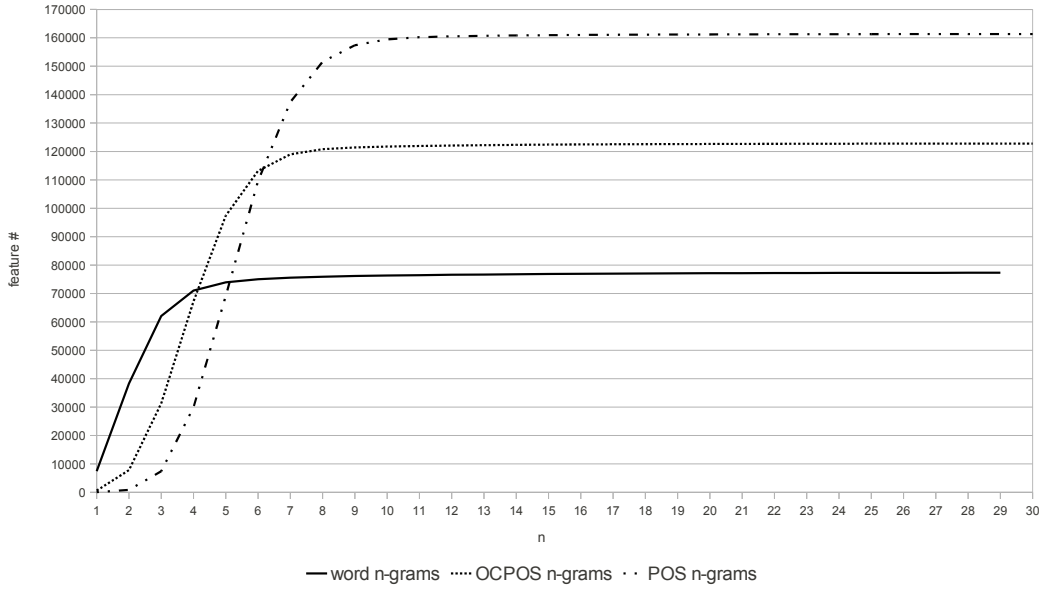

Figure 5.2: Feature counts for $[1, n]$ n-gram settings for the single ICLE sample

There are large differences in terms of feature counts, depending on the particular n-gram class and the value of n used. The figures show that increasing the number of different POS tags leads to more possible different features (up to about 160,000 in our setup). The reason for that is the ability of POS to bridge some break points in the word sequences (i.e., places where different words occur, thus ending a recurrent surface n-gram), and hence to lead to more longer n-grams. Thus the n-grams including POS tags may also reach higher n values: For the word-based n-grams $max_n(d) = 29$, whereas POS-based n-grams reach $max_n(d) = 30$ in our training set.

As expected, the feature counts fall rapidly as the n value passes a certain (n-gram class dependent) threshold (see Figure 5.1). Longer n-grams may potentially contain some specific information not contained in the shorter ones – they may capture, e.g., transliterations of native idioms (Milton & Chowdhury, 1994). So we do not discard any features a priori. The aim is to investigate up to which value of n the n-grams may be worth considering for the given task, despite being rare.

We use binary feature vectors as classifier input, i.e., each essay is represented by a vector containing $\{0, 1\}$ values. If an essays contains a particular n-gram, then the corresponding value in the vector is 1, and 0 otherwise. Since the n-gram frequencies (especially in case of the longer ones) are rather low, we consider such a representation to be a reasonable simplification.

5.2.3 Tools

To extract all recurring n-grams, we implement a dynamic programming algorithm collecting all n-grams of length n based on the n-grams collected for $n - 1$. The algorithm terminates once no n-grams for a given length can be found in the given data. To obtain the n-gram classes incorporating POS tags, we used the *OpenNLP* POS-tagger (`http://opennlp.apache.org`). To choose the classifier to use, we conducted several preliminary tests employing different ML tools. We explored using *TiMBL* (Daelemans et al., 2007), which provides an implementation of the k-NN algorithm, incorporating a range of distance metrics. We then tested different Support Vector Machines (SVMs) which are well-known for their ability to handle large feature sets: *WEKA SMO* (Platt, 1998; Hall et al., 2009),

LIBSVM (Chang & Lin, 2011), and *LIBLINEAR* (Fan et al., 2008). In our trials, the *LIBLINEAR* classifier yielded the best results and was in addition usually faster than the others as well. Hence, we employ the *LIBLINEAR* classifier in our study.

5.2.4 Results

The accuracy for all feature settings is presented in Figures 5.3 and 5.4.

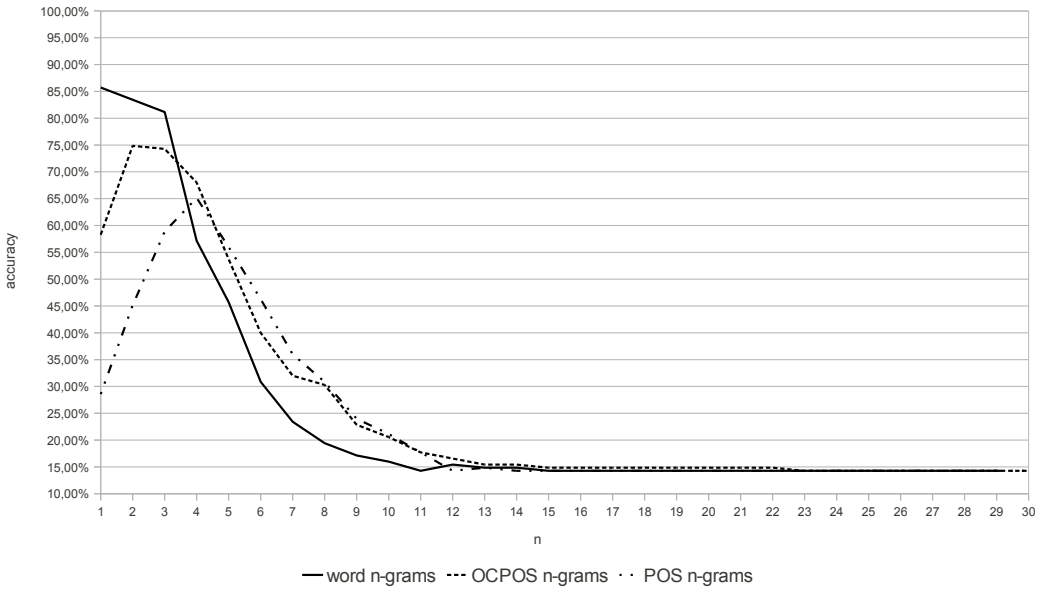

Figure 5.3: Results for single n n-gram settings for the single ICLE sample

Figure 5.3 shows the classification accuracy for all n values of the n-grams separately (i.e., for uni-grams, bi-grams, tri-grams, etc.), whereas Figure 5.4 depicts the classification accuracy for all $[1, n]$ intervals (i.e., for uni-grams alone, uni- and bi-grams together, uni-, bi-, tri-grams together, etc.). There are seven different L1s as classes, each represented by an equal number of essays, so 14.3% is the chance baseline against which to interpret the results.

Best accuracy range The highest accuracy achieved by our recurring n-gram approach is 89,7% using word-based n-grams with intervals from $[1, 2]$ to $[1, 4]$.

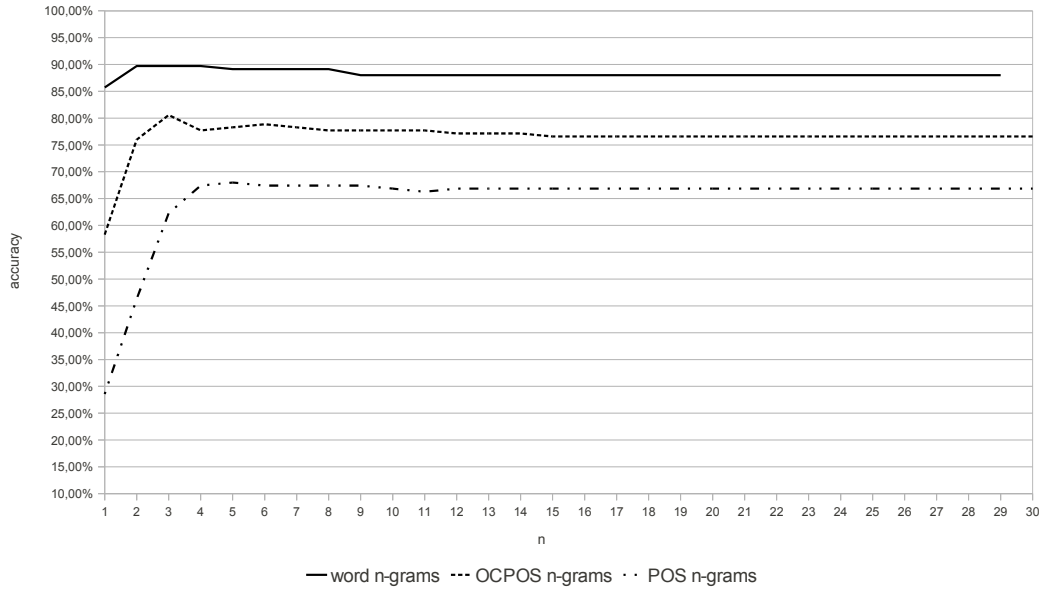

Figure 5.4: Results for $[1, n]$ n-gram settings for the single ICLE sample

This is 16% higher than the best result reported by Wong & Dras (2009) and about 8% higher than that reported by Wong & Dras (2011) on a comparable data set. Brooke & Hirst (2011) reported a slightly better result, 93.8% for seven L1s, but as discussed in Section 3.1 they used more data and a different data split.

	BG	CN	CZ	FR	JP	RU	SP
BG	23	0	0	0	0	2	0
CN	0	24	0	0	1	0	0
CZ	0	0	23	1	0	1	0
FR	1	0	0	22	0	0	2
JP	0	0	1	0	24	0	0
RU	1	0	3	1	1	19	0
SP	1	1	0	1	0	0	22

Table 5.1: Confusion matrix for the best result for the single ICLE sample: 89,7%, word-based n-grams, $[1, 2]$; BG: Bulgarian, CN: Chinese, CZ: Czech, FR: French, JP: Japanese, RU: Russian, SP: Spanish

The confusion matrix in Table 5.1 shows the distribution of correctly classified

as well as misclassified samples for each of the L1s. The performance for the different L1s is generally comparable. Only the result for Russian is slightly below the others.

However, there are clear differences in terms of accuracy between the n-gram classes utilized in this study. As mentioned above, the best result is obtained using pure surface forms, the word-based n-grams. The more different POS tags are incorporated, i.e., the bigger the step from the surface to the more general forms, the lower the accuracy. The information loss involved in the abstraction thus outweighs the broader applicability. The best results are presented in detail in Table 5.2.[4]

	n intervals			single n		
features	$[1, n]$	accuracy	feature #	n	accuracy	feature #
word n-grams	2	89.7%	38,300	1	85.7%	7,446
OCPOS n-grams	3	80.6%	31,263	2	74.9%	7,176
POS n-grams	5	68.0%	69,139	4	65.1%	22,462

Table 5.2: Best results for the single ICLE sample

Table 5.2 shows that POS-based n-grams, i.e., features at the highest generalization level, yield about 13% lower accuracy than the Open-Class-POS-based n-grams, and the latter are performing about 9% worse than word-based n-grams. There is a gap of about 22% between the surface-based and the most generalized n-gram representation used in our study. However, even the most general POS-based n-grams still yield a result of 68%, which is reasonably high considering the baseline of 14.29%. The accuracy of 80.57% obtained using Open-Class-POS-based n-grams is in line with the best results published for a comparable data set.

Different n values Using intervals of n always leads to better results than using n-grams of a particular single n value alone (see Figures 5.3 and 5.4). One can also see that the more POS generalization is incorporated, the longer n-grams are needed to obtain the best results. In this study, the accuracy benefited from n-

[4]If more than one setting per feature class yields the same best accuracy, only the lowest n or $[1, n]$ interval is listed.

grams up to $n = 5$. Thus n-grams with $n > 3$, which are generally not considered in the related research, are not a priori useless.

The longer n-grams in the range of $6 \leq n \leq 10$ seem to be too sparse to improve on the accuracy obtained by intervals of shorter n-grams, at least in the data used in this study. There are a lot of different n-grams in that range, especially for n-grams with POS incorporated (see Figure 5.1), but the impact of lots of different features, with each occurring only in a few essays, seems to be very limited. Moreover, using them in intervals with n-grams of lower n values usually decreases the accuracy (see Figure 5.4). Thus they seem to introduce some noise into the feature set. However, increasing the size of the data set or incorporating more sophisticated generalizations may still allow such n-grams to become useful.

Finally, "very long" n-grams, i.e., n-grams with $n > 10$, usually encode a few, predetermined phrases, such as the wording of the topic the essay is about, or consist of some other copied passages. While such features can be clearly useful for tasks such as plagiarism detection, they are unlikely to be relevant for NLI.

Reliability of the findings Since the results described above are based on a single experiment, one may wonder how generalizable those findings are. As mentioned in Section 5.2.1, we thus conducted nine further experiments. Summing up the results of the ten experiments, we computed the *mean accuracy* values along with the *Sample Standard Deviations (SSD)*. Given that the $max_n(d)$ value varies for the ten training sets, one cannot average over all n for all of the experiments. But as discussed in the previous paragraph, n-grams with $n > 10$ are unlikely to be useful for the purposes of the given task. Hence, we report the accuracy results for the $1 \leq n \leq 10$ range. Figure 5.5 shows the results for $[1, n]$. Overall, the means curves are very similar to the curves we presented in Figure 5.4.

The overall best outcomes are shown in Table 5.3. The best mean accuracy result of 89.4% is yielded by the same setting, namely by the word-based n-grams using the $[1, 2]$ interval.

This best mean accuracy over ten experiments is only 0.3% lower than the corresponding best result from the single experiment described in the *Best accuracy range* paragraph above (see p. 47). The SSD with values around 2% for the best performing settings indicates little variance among the experiments.

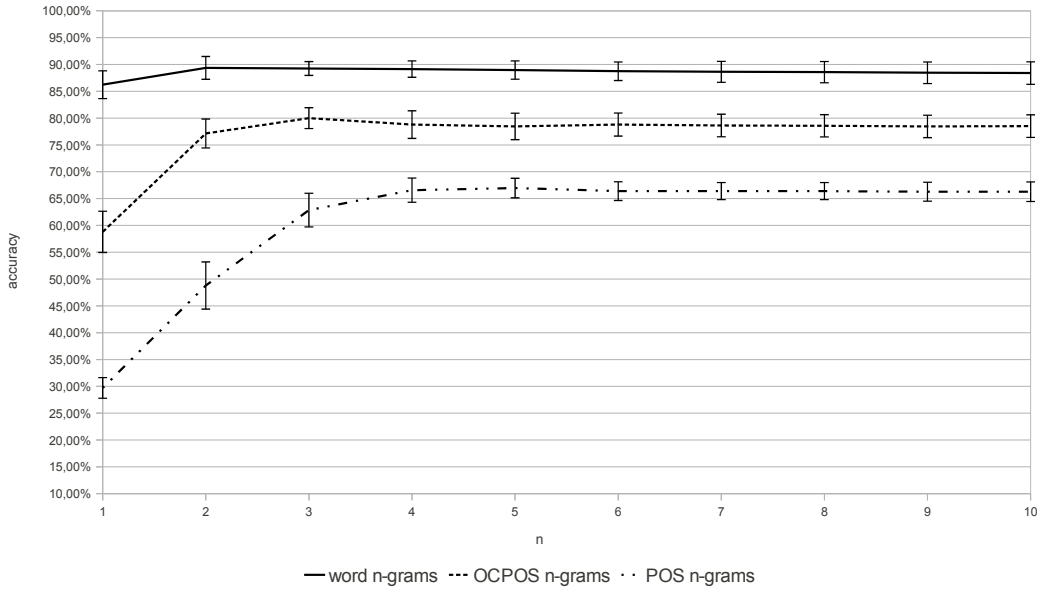

Figure 5.5: *Mean accuracy* and *SSD* for $[1, n]$ n-gram settings for the ten ICLE samples

	n intervals			single n		
features	$[1, n]$	mean accuracy	SSD	n	mean accuracy	SSD
word n-grams	2	89.4%	2.1%	1	86.2%	2.6%
OCPOS n-grams	3	80.0%	1.9%	2	73.7%	2.7%
POS n-grams	5	67.0%	1.8%	4	60.9%	3.4%

Table 5.3: Best mean accuracy results for ten ICLE samples

Discussion The ICLE contains essays from a range of topics, so one may wonder about the impact of the contents on the NLI task. Using only essays of the same topic would in principle be preferable, but it would significantly reduce the amount of data available. As mentioned in Section 3.1, Tsur & Rappoport (2007) argued that such a content bias is rather marginal for the subset of the ICLE they used. In contrast, the findings of Brooke & Hirst (2011) suggested a high topic bias in the ICLE data. In order to obtain more independence from the content of an essay, there is a clear need for some abstraction away from the surface encoding form and meaning together. Yet, the features in our study with the highest level of generalization and thus probably the lowest topic bias, recurring POS-

based n-grams, provide results about 22% below those purely based on surface forms. A combination of surface and generalized forms may be a reasonable middle ground. In that light, the *Open-Class-POS-based n-grams* appear attractive since they replace many of the topic-specific meaning distinctions with POS-tags. They are less tied to the meaning than word-based n-grams, but still yield high accuracy with relatively low feature counts in the best performing n range. At the same time, Brooke & Hirst (2011) observe a comparable drop for word and POS-based features in cross-corpus evaluation with the Lang-8 corpus, and Golcher & Reznicek (2011) show that POS n-grams still contain information relevant to topic classification for the German learner corpus FALKO. More research thus is needed to verify which features are sufficiently general and applicable across corpora. We address this issue in the next section.

5.3 Cross-corpus Generalizability of the Findings

To address the question whether the models trained and evaluated on the ICLE corpus generalize to other learner corpora, we conducted a second study.

5.3.1 Data

In this second study, we use four different learner corpora. Complementing the ICLE introduced above, we use the NOCE, USE and HKUST corpora compiled by independent research teams.

NOCE: The Non-Native Corpus of English (Díaz Negrillo, 2007, 2009) This is an English learner corpus consisting of mainly argumentative essays on several topics written by L1 Spanish speakers. The data was collected at the University of Granada and the University of Jaén using texts by undergraduates pursuing an English degree. The corpus contains 1,022 essays.

USE: The Uppsala Student English Corpus (Axelsson, 2000, 2003) This is a corpus of learner English consisting of texts written by Swedish students at the Department of English at Uppsala University. The texts contained in the corpus

are essays written as part of the regular curriculum and cover several topics of different genres, e.g., argumentation, reflection, literature course assignment, etc. The corpus contains 1,489 essays. Since the essays from the other corpora used in this study are mostly argumentative, to obtain comparable data in terms of the text properties we use only the argumentative subset of the corpus (from the first term). This subcorpus consists of 344 essays.

HKUST: The Hong Kong University of Science and Technology English Examination Corpus (Milton & Chowdhury, 1994) This is an English learner corpus containing texts written by L1 Chinese speakers. The version of the corpus we are using consists of 1,100 argumentative essays on different topics collected 1992 during the public matriculation examination, which is taken each year by students leaving secondary school. For the present work, we took a 8% random sample of the whole corpus, consisting of manually tagged 77 essays as described in Milton & Chowdhury (1994).

Preprocessing and data setup As preprocessing, we removed all types of meta-information and annotation contained in the learner corpora (personal information about the author of the text such as the age or the L1, topic tags, error annotation, etc.) as well as all punctuation marks, special characters and capitalization, and we tokenized the essays. Hence, as in the first study each text is represented as an array of lower-case words.

Based on the data described above, we explore the NLI task using a setup with three L1s: Spanish, Swedish and Chinese. First, we compile *two separate test sets*. The first test set consists of randomly selected 70 essays per L1 from ICLE. To compile the second test set, we randomly select 70 essays per L1 correspondingly from HKUST and USE and 140 essays from the NOCE corpus. Since the NOCE essays tend to be shorter than the other ones, we merge the 140 essays pairwise to obtain 70 texts of a size comparable to the essay size from the other corpora. The texts on average contain 620 words. Second, we compile *ten separate training sets*. Each training set consists of randomly selected 140 essays per L1 from the overall ICLE corpus (without the essays selected for the ICLE test). Thus we obtain ten separate training sets with 420 essays each, randomly selected from the

ICLE corpus; and two separate test sets with 210 texts each, one compiled using ICLE alone and another compiled using NOCE, USE and HKUST.

This setup allows us to perform ten *single-corpus* evaluations (i.e., training and testing on the same corpus) on the ICLE data alone, as well as ten *cross-corpus* evaluations (i.e., training on the one corpus and testing on another corpus) using ICLE data for training and NOCE, USE, HKUST data for testing. With ten separate ICLE training sets, we are able to build ten different classifier models and to observe the variance in the generalizability of the patterns learned on different ICLE subsets. We thus are able to observe the generalizability of the ICLE patterns to other corpora in direct comparison to ICLE itself.

5.3.2 Results

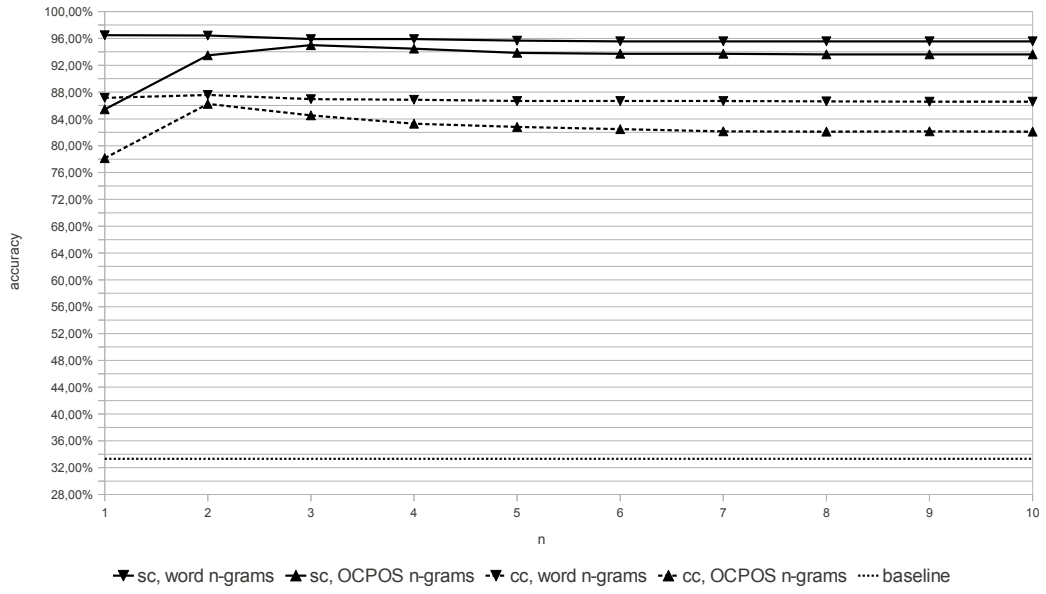

Figure 5.6: *Mean accuracy* for $[1, n]$ n-gram settings for the ten ICLE training sets (sc = single-corpus, cc = cross-corpus evaluation)

Based on the ten different training sets, we conducted tests for each $[1, n]$ n-gram interval with $1 \leq n \leq 10$ using the two best performing n-gram classes (i.e., word- and OCPOS-based n-grams as features), and performed both a single-

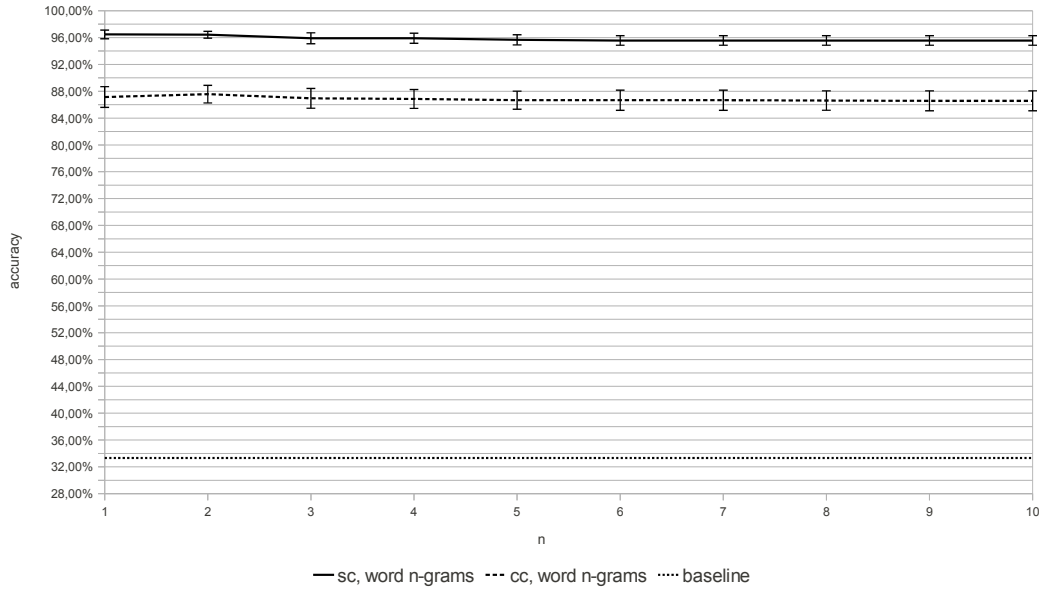

Figure 5.7: *Mean accuracy* and *SSD* for $[1, n]$ n-gram settings for the ten ICLE training sets, recurring *word-based n-grams* as features (sc = single-corpus, cc = cross-corpus evaluation)

corpus evaluation and a cross-corpus evaluation. We thus obtained 400 separate accuracy values overall (10 training sets · 2 n-gram classes · 10 n-gram intervals · 2 evaluation types).

Figure 5.6 sums up the results by depicting the *mean accuracy* values on the two test sets, obtained using ten different training sets for both n-gram classes and each of the ten n-gram intervals, along with the chance baseline. Since in this set of experiments we employ three different L1s, each represented by an equal number of essays, the chance baseline is 33.3%.

We left the SSD bars out of Figure 5.6 to keep it readable, but it naturally is interesting to consider the variance. Figure 5.7 shows the single- and cross-corpus accuracies for the word-based n-grams from Figure 5.6 together with the corresponding SSD. Figure 5.8 presents the same for the OCPOS-based n-grams. In both figures the variance is low; as expected, the cross-corpus evaluation shows slightly higher SSD values.

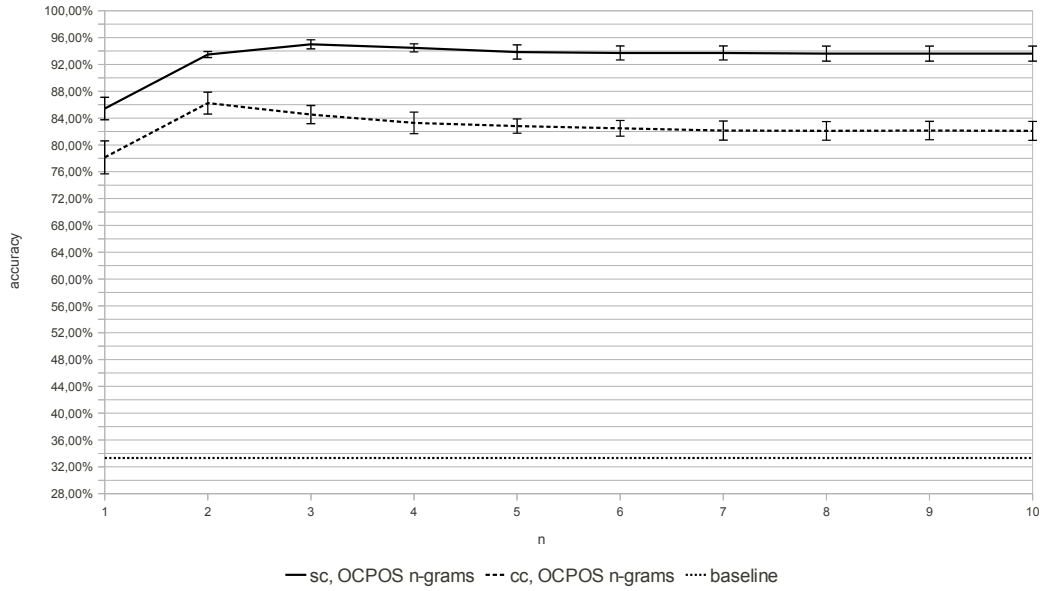

Figure 5.8: *Mean accuracy* and *SSD* for $[1, n]$ n-gram settings for the ten ICLE training sets, recurring *OCPOS-based n-grams* as features (sc = single-corpus, cc = cross-corpus evaluation)

features	evaluation	$[1, n]$	mean accuracy	SSD
word n-grams	single-corpus	1	96.5%	0.6%
	cross-corpus	2	87.6%	1.3%
OCPOS n-grams	single-corpus	3	95.0%	0.7%
	cross-corpus	2	86.2%	1.6%

Table 5.4: Best results for ten ICLE training sets

Table 5.4 shows the best accuracies for both feature classes along with the corresponding SSD values obtained on the two different evaluation types as well as the corresponding n intervals. Though the best performing n-gram intervals differ for both feature classes on single-corpus evaluation, in the cross-corpus evaluation *recurring bi-grams* perform best for both.

At the end of Section 5.2.4, we hypothesized that the more abstract OCPOS-based n-grams may perform better than the surface-near word-based ones in cross-corpus evaluation. However, the accuracies obtained using word-based n-grams

are on average as good or better than the ones obtained using OCPOS-based n-grams (see Figure 5.6 and Table 5.4). Apparently learners with different L1s make lexical choices which are indicative across a range of topics. A first qualitative analysis using *SvmAttributeEval* and *InfoGain* feature selection techniques via *WEKA* (Hall et al., 2009), points to the use of predicates such as *get*, *take*, *choose*, *make use of*, *consider*, *be able to*, *understand*, or *suggest*. A precise characterization of the nature of this lexical material seems relevant to investigate in future work.

Domain dependence The experiments we ran with the NOCE, USE and HKUST corpora show far higher accuracies for the cross-corpus evaluation than what is reported by Brooke & Hirst (2011) for the Lang-8 corpus. In a setup with a chance baseline of 14.3%, Brooke & Hirst (2011) report 70.1% – 93.8% (depending on the employed feature set) on single-corpus evaluation using ICLE, but only 22.0% – 29.3% for cross-corpus evaluation, training on ICLE and testing on Lang-8.[5] In contrast, in a setup with a chance baseline of 33.3% we obtained a best result of 95% – 96.5% (depending on the employed n-gram class) on single-corpus evaluation using ICLE, and 86.2% – 87.6% in a cross-corpus evaluation setup with training on ICLE and testing on NOCE/USE/HKUST (see Table 5.4 and Figure 5.6). Thus when using ICLE for training and another corpus instead of ICLE for testing, there is a drop of about 64% in Brooke & Hirst (2011) but only around 9% in our work. Such a big discrepancy can be hardly explained by the given difference in the baselines alone. Brooke & Hirst (2011) argue that the observed drop in accuracy seems to be due to a topic bias in ICLE, allowing for easy classification by topic instead of by L1 in single-corpus experiments (see Section 3.1).

The corpora we used for the cross-corpus evaluation were compiled by different research teams using own essay topic lists. To investigate whether there still may be some topic overlap contributing to the promising classification outcomes

[5]In the first version of the paper Brooke & Hirst (2011) provided online by the authors, and used for our discussion in Bykh & Meurers (2012) – the contribution this chapter is mostly based on –, Brooke & Hirst reported cross-corpus accuracies of 15.7% – 17.0% instead of 22.0% – 29.3% for training on ICLE and testing on Lang-8. According to the authors, the first, worse results, emerged due to an evaluation error, and the figures were updated later on. However, the improved performance does not change the main point and the conclusions in this paragraph.

in our experiments, we extracted the topics from our NOCE/USE/HKUST test set as well as from the ICLE training set yielding the best cross-corpus evaluation results. In both cases there were more than 100 different topics, and none of them matched between ICLE used for training and NOCE/USE/HKUST used for testing in the cross-corpus setup. Thus topic overlaps seem very unlikely to have notably skewed the results in our cross-corpus evaluation.

On the one hand, topic-related issues can still play a role in evaluations using ICLE. We also encountered a drop in accuracy in our cross-corpus experiments after all. On the other hand, our findings suggest that some of the notable performance decrease observed in Brooke & Hirst (2011) might also occur due to certain characteristics of the Lang-8 corpus, rather than to a general failure of the models learned on the ICLE to generalize to other learner corpora. The lack of consistency in the Lang-8 pieces combined into documents, or the very different nature of the ICLE and the Lang-8 data with respect to the general structure and the genre, seems to play a role as well.

5.4 Further Exploring Linguistic Generalization

The findings in this chapter obtained so far suggest that features, such as the POS- and OCPOS-based n-grams, incorporating some linguistic generalization, do not provide a quantitative edge over the pure surface-based n-grams. In this section, we further investigate the question, whether linguistic generalization can make a quantitative difference in the context of n-grams. For this, we propose two new feature versions based on recurring n-grams, namely, *POS generalized* (*POS-gen*) and *freely generalized* (*Free-gen*) *word-based n-grams*. It is important to note that both use word-based n-grams as basis. The difference is that the former feature version incorporates linguistic abstraction from words to POS, whereas the latter is based on an abstraction not incorporating linguistic knowledge. For our experiments we used the basic ICLE data setup introduces in Section 5.2.1. This section reports some of our results published in Meurers, Krivanek & Bykh (2014).

Feature generation We employ all the recurring word-based n-grams with $2 \leq n \leq 6$ obtained from the ICLE training set. The results presented above suggest that longer n-grams are unlikely to be effective as features, and we do not include uni-grams in this experiment because the point here is to compare recurring n-grams with the POS-gen and Free-gen generalizations, which only come into play for longer n-grams. We start with the bi-grams, which are the same for all three feature types and can serve as a baseline.

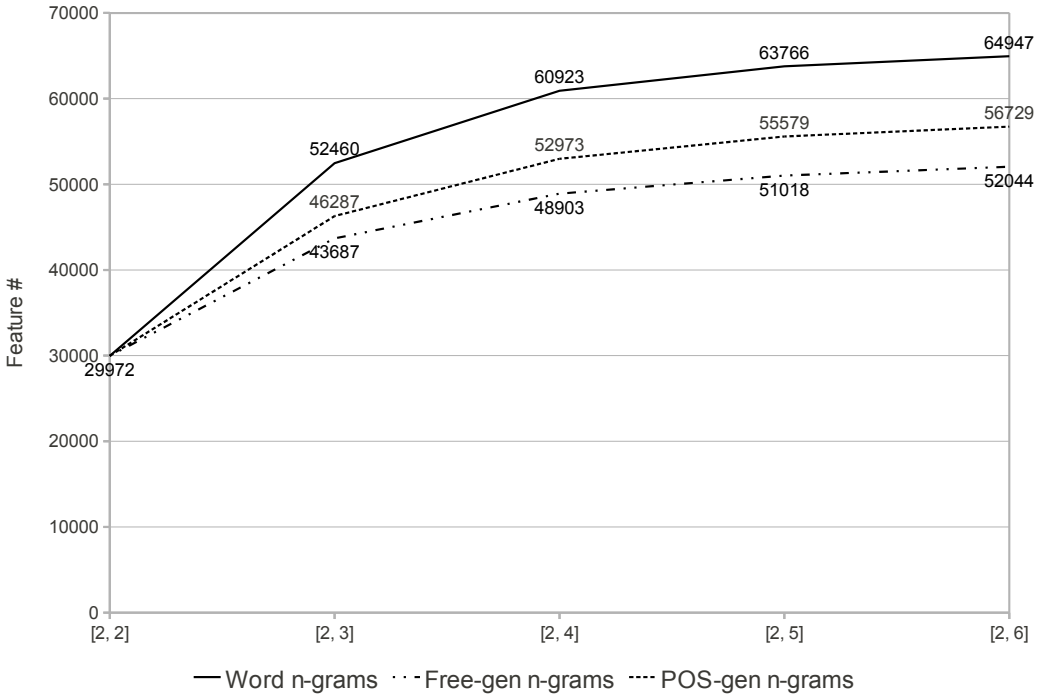

Figure 5.9: Feature counts for generalized word-based n-grams

Based on that extracted feature set, we generate the POS-gen and Free-gen features in the following way: For all word-based n-grams with $n \geq 3$ we retain the words at the boundaries of the n-grams and replace the rest, i.e., the middle part of the n-grams, by the corresponding POS-tags for the POS-gen features, and by a wildcard (*) for the Free-gen features. Figure 5.10 shows an example for n-grams of length five.

Recurring word-based n-grams:	*He gave John the book*
Recurring POS-gen word-based n-grams:	*He VBD NNP DT book*
Recurring Free-gen word-based n-grams:	*He * * * book*

Figure 5.10: Example for generalized word-based n-grams with $n = 5$

While these three representation options allow us to investigate the impact of a linguistic and a non-linguistic abstraction away from the surface n-gram, one should keep in mind that they are just two of the many abstraction options one could investigate. Moreover, we here generalize the word-based recurring n-grams we obtained from the corpus; another option would be to start by collecting all POS-based n-grams recurring in the corpus and to use these as basis, etc.

The results in Section 5.2 show that intervals of n-grams outperform single n n-grams, thus we consider intervals of n-grams. In particular, we employ intervals from $[2, 2]$ to $[2, 6]$. For the ICLE training set used here, we obtain the feature counts for the different intervals shown in Figure 5.9. Depending on the feature version and the interval used, we obtained up to 65,000 features. The more abstract the features become, from words-based via POS-gen to Free-gen, the fewer features there are, given that a single more general feature often subsumes multiple specific ones. We use a binary feature representation for our classification experiments in this section (see Section 5.2.2).

Results for the generalized word-based n-grams Figure 5.11 presents the classification accuracy for the different n intervals of n-grams on the held out ICLE test set. Using bi-grams alone yields an accuracy of 86.9%, which is the same for all three feature types since in the abstractions explored here the n-gram boundaries always consist of words. The best accuracy for the recurring word-based n-grams as our baseline feature type is 87.4% using the interval $[2, 4]$. The interesting question now is: What is the effect of the two different types of abstraction, i.e., the linguistic (via POS) and the non-linguistic one (via *)? As shown in the Figure 5.11, incorporating non-linguistic abstraction (Free-gen), decreases the classification accuracy. The words in the middle of the n-gram thus include information relevant for distinguishing the different native languages. On

the other hand, incorporating some linguistic abstraction using the POS-gen n-grams increased the classification accuracy. The best result for the POS-gen n-grams is 88.6% for the interval $[2, 4]$, which is the best result in this section. In other words, incorporating some linguistic abstraction via POS and including sequences longer than the commonly used tri-grams is most successful in providing access to information relevant for NLI.

Systems based on the generalized word-based n-grams explored here, did not outperform those based on the pure word-based n-grams in the interval $[1, 2]$ (see Section 5.2.4). Nevertheless, the findings show that generalized word-based n-grams incorporating linguistic abstraction can indeed provide a quantitative edge over non-linguistic generalizations on n-grams, or pure word-based n-grams of the same length n – Especially, if the generalization is applied to the conceptually more interesting n-grams of higher n, capturing more structural information.

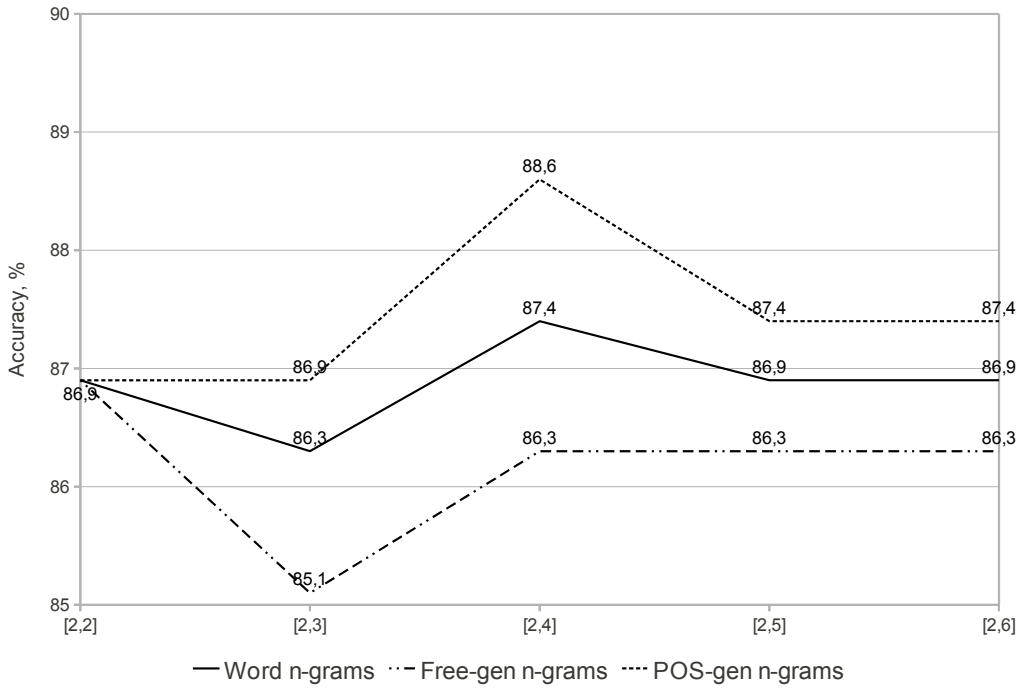

Figure 5.11: Accuracy for the generalized word-based n-grams

5.5 Conclusions

In this chapter, we explored using three different classes of *recurring n-grams* as features for NLI, namely, *word-*, *POS-* and *Open-Class-POS-based n-grams*. We used these features in a ML setup employing a Support Vector Machine (SVM) classifier on randomly selected data from the ICLE corpus incorporating seven different L1s. The best performing class are the word-based n-grams with an accuracy of 89.7%, which compares well to the 81.7% reported by Wong & Dras (2011) as the highest accuracy achieved in the previous work[6] on comparable data.

To investigate the variance, we conducted nine further experiments based on random samples from ICLE. The mean accuracy values obtained from the overall ten experiments are very similar to those from the first experiment. The variance of the outcomes is moderate, with SSD around 2% for the best performing settings. The bigger the step from the surface-based to more generalized features, the lower the accuracy. The recurring n-gram approach employing Open-Class-POS-based n-grams yields an accuracy of 80.6% and using POS-based n-grams we obtained 68%, which still is reasonably high considering the chance baseline of 14.3% for this task.

We then investigated the claim in Brooke & Hirst (2011) that models trained on the ICLE corpus do not generalize to other learner corpora. For this purpose we conducted a second set of experiments comparing *single-corpus* and *cross-corpus* results. In contrast to their cross-corpus findings using the Lang-8 corpus, our results show that the patterns learned on ICLE can generalize to other learner corpora. In particular, we showed that training on ICLE and testing on three independently collected corpora, namely, NOCE, USE and HKUST, still yields reasonably high accuracies of about 88% for a NLI classification task with three L1s. The low results for the Lang-8 corpus reported in Brooke & Hirst (2011), might be partially due to the lack of consistency in the Lang-8 pieces combined into documents, or the very different nature of the ICLE and the Lang-8 data with respect to the general structure and genre.

Finally, we showed that linguistic abstraction applied to pure recurring word-

[6]By "previous work" we here refer to the results published before Bykh & Meurers (2012), the contribution this chapter is mainly based on. For a discussion on subsequent relevant contributions see Section 3.2 and Section 3.3.

based n-grams can indeed provide quantitative advantages, especially if applied to n-grams of higher n.

Connecting the findings and the research questions Regarding the particular five research questions in the focus of this thesis (see Section 1.3), the findings in this chapter contribute to three of them as follows:

1. [LINGUISTIC-FEATURES]: In this chapter, we started our explorations on NLI with a systematic investigation of surface features, namely, recurring word-based n-grams. This was not done to that extent before. Next, we applied different levels of linguistic abstraction to the n-grams by incorporating POS information. It turned out that the pure surface-based n-grams outperformed the more abstract features in both, single-corpus and cross-corpus settings. In general, word-based n-grams seem to constitute one of the strongest baselines in NLI. This is an insight, supported by several contributions which were concurrent to Bykh & Meurers (2012) or published later (Tetreault et al., 2012; Brooke & Hirst, 2012b, etc., see Chapter 3 and Chapter 14). In particular, this was also confirmed by the results of the First NLI Shared Task (Tetreault et al., 2013) as discussed in Section 3.2. However, combining surface and some more abstract features might show some performance improvements, a question which we will mainly target in Chapter 6 and Chapter 14.

2. [GENERAL-STRUCTURES]: Compared to the pure surface-based n-grams, the more abstract n-grams incorporating POS tags are better capable of capturing general linguistic structures, in that such features abstract from individual words to the more general POS classes. Some recent research shows that related features can provide interesting linguistic insights (Malmasi & Dras, 2014d). Furthermore, n-grams incorporating POS are more frequent and useful up to higher n levels. That makes them promising candidates for settings suffering from high data sparsity. Finally, we showed that using recurring word-based n-grams as basis and introducing a linguistic generalization via POS to parts of them, can indeed provide better results compared to the corresponding versions without generalization or with non-linguistic

generalization – Especially, if the generalization is applied to the conceptually more interesting n-grams of higher n capturing wider structural information.

3. [CROSS-CORPUS]: Different from the findings in Brooke & Hirst (2011), our cross-corpus results show a rather moderate drop in accuracy compared to the single-corpus settings, suggesting that the patterns learned by the classifier on ICLE can generalize across data sets. One of the reasons for this discrepancy might be that the NOCE/USE/HKUST data used in our cross-corpus experiments, is better comparable to ICLE in terms of the genre and the general structure of the documents. Nevertheless, the cross-corpus results are clearly below the single-corpus outcomes, pointing to one of the general challenges in the given research area (see Section 3.3.1 and Chapter 6). Finally, our findings suggest that lexical features, such as word-based n-grams seem to play an important role for the cross-corpus performance in NLI (see also Chapter 6 and Chapter 14, cf. Brooke & Hirst, 2012b).

Chapter 6

Exploring Linguistic Features and Feature Combinations

6.1 Introduction

In Chapter 5, which reflects our work in Bykh & Meurers (2012), we explored a data-driven approach employing three different types of recurring n-grams as features: one purely surface-based type and two types incorporating some linguistic abstraction, namely some POS information. In this chapter, we extend our feature space by including linguistic features based on dependency and constituency trees, as well as features encoding some morphological properties, the nature of the realizations of particular lemmas, and several measures of linguistic complexity, which were originally developed for proficiency and readability classification. Our main goals are the following:

1. Implementing some linguistic features, novel for the task of NLI, and investigating their performance in particular.

2. Investigating the performance of systems which are based on a broad feature set, including different types of novel and previously used features.

    - Exploring the joint performance of all features used for this study.
    - Exploring the performance of different systems based on various feature combinations to get more insight about the employed feature set.

This study was conducted in the context of the First NLI Shared Task (see Section 3.2), and most of the findings discussed in this chapter were reported in Bykh et al. (2013).[1] Our id in the shared task was *TUE*.

6.2 Tasks

The study presented in this chapter, follows the setup of the First NLI Shared Task as described in Section 3.2.2. It contributes to each of the three tasks, employing the same TOEFL11 *test* set for *testing*, and differing in the *training* data as follows:

1. *Closed* task: TOEFL11 *train*, TOEFL11 *dev*
2. *Open-1* task: any data, excluding: TOEFL11 *train*, TOEFL11 *dev*
3. *Open-2* task: any data

Technically, depending on the actual task, there are *single-corpus* (*closed* task), *cross-corpus* (*open-1* task) or *"extended" single-corpus* (*open-2* task) settings.

We call the last setting *"extended" single-corpus* for the following reason: The training set in the *open-2* task consists of TOEFL11 and other corpora, and the testing is performed on the TOEFL11 *test* set, which means that we have neither a single-, nor a cross-corpus setting in the strict sense. From our point of view, such a setting is theoretically closer to the single-corpus case, namely, we consider it a special case of a *single-corpus* setting, where the training set is *extended* by other corpora.

6.3 Data

In this section, we present the data used for our single- and cross-corpus explorations, discussed in this chapter.

[1]The underlying system was developed under the supervision of *Prof. Dr. Detmar Meurers*, by *Sowmya Vajjala*, *Julia Krivanek* and the *author of this thesis*, who was the coordinator of the team and the lead author of the contribution.

T11: TOEFL11 alias The ETS Corpus of Non-Native Written English (Blanchard et al., 2013) This is the main corpus of the First NLI Shared Task, which was already described in detail in Section 3.2.1, where we discussed the contribution of this competition. In sum, it consists of essays written by English learners with 11 L1s, namely, Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu and Turkish, and from three different proficiency levels, i.e., low, medium and high. Each L1 is represented by a set of 1100 essays (*train*: 900, *dev*: 100, *test*: 100). The labels for the *train* and *dev* sets were given from the start, whereas the labels for the *test* set were provided after the results were submitted to the shared task system.

ICLE: International Corpus of Learner English, second version (Granger et al., 2009) The ICLE corpus was one of the first data sources used for NLI, and it was already introduced in Section 3.1. It consists of 6,085 essays written by English learners of 16 different L1s. They are at a similar level of English proficiency, namely higher intermediate to advanced and of about the same age. For the *open* tasks, we utilized the essays by learners of the seven L1s in the intersection with T11, i.e., Chinese (982 essays), French (311), German (431), Italian (391), Japanese (366), Spanish (248), and Turkish (276).

FCE: First Certificate in English Corpus (Yannakoudakis et al., 2011) The FCE dataset consists of 1,238 scripts produced by learners taking the First Certificate in English exam, assessing English at an upper-intermediate level. For the *open* tasks, we used the essays by learners of the eight L1s in the intersection with T11, i.e., Chinese (66 essays), French (145), German (69), Italian (76), Japanese (81), Korean (84), Spanish (198), and Turkish (73).

BALC: BUiD (British University in Dubai) Arab Learner Corpus (Randall & Groom, 2009) The BALC corpus consists of 1,865 English texts written by L1 Arabic learners from the last year of secondary school and the first year of university. The texts were scored and assigned to six proficiency levels. For the *open* tasks, we utilized the data from the middle proficiency range, i.e., from the levels 3–5, amounting to overall 846 texts.

ICNALE: International Corpus Network of Asian Learners of English (Ishikawa, 2011) The utilized version of the ICNALE corpus consists of 5,600 essays written by college students in ten countries and areas in Asia as well as by English native speakers. The learner essays are assigned to four proficiency levels following the CEFR guidelines (A2, B1, B2, B2+). For the *open* tasks, we used the essays written by learners from Korea (600 essays) and from Pakistan (400).[2] Without access to a corpus with Hindi as L1, we decided to label the essays written by Pakistani students as Hindi. Most of the languages spoken in Pakistan, including the official language Urdu, belong to the same Indo-Aryan/-Iranian language family as Hindi. Our main focus here was on avoiding overlap with Telugu, the other Indian language in this shared task which belongs to the Dravidian language family.

TÜTEL-NLI: Tübingen Telugu NLI Corpus We collected 200 English texts written by Telugu native speakers from bilingual (English-Telugu) blogs, literary articles, news and movie review websites.

NT11: NON-TOEFL11 We combined the ICLE, FCE, ICNALE, BALC and TÜTEL-NLI sources discussed above, in the *NT11* corpus, consisting of overall 5,843 essays for 11 L1s, as shown in Table 6.1.

6.4 Features

In this section, we provide a description of the feature set used for the study in this chapter. We explore a wide range of features for developing our model. Some of the features, such as *recurring function based dependencies* or the different *suffix-based features*, are novel for the task NLI. Other features were already used in the related work.

For each feature, we define short identifiers, such as *rc. word ng.*, *dep. pos* or *suffix bin.*, etc. For convenience, we will use such identifiers to refer to the corresponding features in the various tables, as well as in the following explanations and discussions.

[2]We did not include ICNALE data for more L1s to avoid over-representation of already well-represented Asian L1s.

L1	corpora					#
	ICLE	FCE	BALC	ICNALE	TÜTEL	
ARA	-	-	846	-	-	846
CHI	982	66	-	-	-	1048
FRE	311	145	-	-	-	456
GER	431	69	-	-	-	500
HIN	-	-	-	400	-	400
ITA	391	76	-	-	-	467
JPN	366	81	-	-	-	447
KOR	-	84	-	600	-	684
SPA	248	198	-	-	-	446
TEL	-	-	-	-	200	200
TUR	276	73	-	-	-	349
#	3005	792	846	1000	200	5843

Table 6.1: Distribution of essays for the 11 L1s in NT11

6.4.1 Recurring N-grams

Following our approach in Chapter 5, we used recurring word-based n-grams (*rc. word ng.*), i.e., all word-based n-grams occurring in at least two texts of the training set. We focused on recurring uni-grams and bi-grams, which in our previous work and in T11 testing with the *dev* set worked best. As in our previous work, we used a binary feature representation encoding the presence or absence of the n-gram in a given essay.

In addition, we employed recurring OCPOS n-grams (*rc. OCPOS ng.*), i.e., all OCPOS n-grams occurring in at least two texts of the training set, which were obtained as described in Chapter 5. OCPOS means that the open class words (nouns, verbs, adjectives and cardinal numbers) are replaced by the corresponding POS tags. For POS tagging we used the *OpenNLP* toolkit[3]. In our previous work, recurring OCPOS n-grams up to length three performed best (see Chapter 5). However, for T11 we found that including four- and five-grams was beneficial. This confirms our assumption that longer n-grams can be sufficiently common to be useful (see Section 5.2.4). Thus, we used the recurring OCPOS n-grams up

[3] `http://opennlp.apache.org`

to length five for the experiments in this chapter. We again employed a binary feature representation.

6.4.2 Dependency

Stanford dependencies Tetreault et al. (2012) explored the utility of basic dependencies as features for NLI. In our approach, we extracted all Stanford dependencies (de Marneffe et al., 2006) using the trees assigned by the *Berkeley Parser* (Petrov & Klein, 2007). We considered lemmatized typed dependencies (*type dep. lm.*) such as *nsubj(work,human)*, as well as POS-tagged ones (*type dep. POS*) such as *nsubj(VB,NN)* for our features. We used count-based features for those typed dependencies.

Recurring MATE dependencies Extending the perspective on recurring pieces of data to other data types, we explored a novel feature type: recurring word-based dependencies (*rc. word dep.*). A feature of this type consists of a head and all its immediate dependents. The dependencies were obtained using the *MATE* parser (Bohnet, 2010). The words in each n-tuple are recorded in lowercase and listed in the order in which they occur in the text; heads thus are not singled out in this encoding. For example, the sentence *John gave Mary an interesting book* yields the following two features *(john, gave, mary, book)* and *(an, interesting, book)*. As with recurring n-grams we utilized only features occurring in at least two texts of the training set, and we used a binary feature representation.

Recurring function-based dependencies (*rc. func. dep.*) constitute another novel feature type. It is a variant of the recurring word-based dependencies described above, where each dependent is represented by its grammatical function. The above example sentence thus yields the two features *(sbj, gave, obj, obj)* and *(nmod, nmod, book)*.

MATE dependencies realization We employ three feature types novel for NLI, capturing some properties of the dependencies, i.e., the number of the dependants and the way the dependents of a particular verb lemma are realized with respect to their POS-tags.

First, we encoded the number of dependents realized by a verb lemma, normalized by the frequency of this lemma. We refer to these features as *Dependency number* (*dep. num.*). For example, if the lemma *take* occurred ten times in a document, three times with two dependents and seven times with three dependents, we get the features *take:2-dependents = 3/10* and *take:3-dependents = 7/10*.

Second, we encode the different dependent-POS combinations for a verb lemma, normalizing the counts by the frequency of this lemma. We refer to these features as *Dependency variability* (*dep. var.*). If in the example above, the lemma *take* occurred three times with two dependents JJ-NN, two times with three dependents JJ-NN-VB, and five times with three dependents NN-NN-VB, we obtain *take:JJ-NN = 3/10*, *take:JJ-NN-VB = 2/10*, and *take:NN-NN-VB = 5/10*.

Third, we employ another feature which is derived from *dep. var.* and encode how frequent which kind of category was a dependent for a given verb lemma. We refer to these features as *Dependency POS* (*dep. POS*). Continuing the example above, *take* takes dependents of three different categories: JJ, NN and VB. For each category, we create a feature, the value of which is the category count divided by the number of dependents of the given lemma, normalized by the frequency of this lemma in the given document. In the example, we obtain *take:JJ = (1/2 + 1/3)/10, take:NN = (1/2 + 1/3 + 2/3)/10*, and *take:VB = (1/3 + 1/3)/10*.

6.4.3 Constituency

Based on the syntactic trees assigned by the *Berkeley* parser (Petrov & Klein, 2007), we extracted all local trees, i.e., trees of depth one and used them as features (*local trees*). This corresponds to using CFG production rules as features (Wong & Dras, 2011). For example, for the sentence *I have a tree*, the parser output is: *(ROOT (S (NP (PRP I)) (VP (VBP have) (NP (DT a) (NN tree))) (. .)))* for which the local trees are *(S NP VP .), (NP PRP), (NP DT NN), (VP VBP NP), (ROOT S)*. We used the count-based feature representation.

6.4.4 Morphology

The use of different derivational and inflectional suffixes may contain information indicative for the particular L1s – either in connection with L1-transfer, or in terms

of what suffixes are taught, e.g., for nominalization. To the best of our knowledge, corresponding morphological features were not used in the previous research on NLI. Here, we implemented a first version of such features as described below.

In a very basic approximation of morphological analysis, we used the porter stemmer implementation of *MorphAdorner*[4]. For each word in a learner text, we removed the stem it identified from the word, and if a suffix remained, we matched it against the *Wiktionary list of English suffixes*[5]. For each valid suffix identified in a text, we defined a binary feature (*suffix bin.*) recording the presence/absence, as well as a corresponding count-based feature (*suffix cnt.*).

We also wondered whether the subset of morphologically complex uni-grams may be more indicative than considering all uni-grams as features. As a simple approximation of this idea, we used the stemmer plus the suffix-list approach mentioned above, and utilized all words containing a suffix as features. We employed both, a binary (*stemsuffix bin.*) and a corresponding count-based (*stemsuffix cnt.*) encodings.

6.4.5 Complexity

We used a wide range of features related to language complexity (*complexity*), with most of them being novel for the task of NLI. Given that the proficiency level of a learner was shown to play a role in NLI (Tetreault et al., 2012), we implemented all the text complexity features from Vajjala & Meurers (2012), who used measures of learner language complexity from SLA research for readability classification. These features consist of lexical richness and syntactic complexity measures from SLA research (Lu, 2010, 2012), as well as other syntactic parse tree properties and traditionally used readability formulae. The parse trees were built using the *Berkeley* parser (Petrov & Klein, 2007) and the syntactic complexity measures were estimated using the *Tregex* package (Levy & Andrew, 2006).

In addition, we included morphological and POS features from the *CELEX* Lexical Database (Baayen et al., 1995). The morphological properties of words in CELEX include information about the derivational, inflectional and compo-

[4] `http://morphadorner.northwestern.edu`
[5] `http://en.wiktionary.org/wiki/Appendix:Suffixes:English`

sitional features of the words along with information about their morphological origins and complexity. POS properties of the words in CELEX describe the various attributes of a word depending on its parts of speech. We included all the non-frequency based and non-word-string attributes from the English Morphology Lemma (EML) and English Syntax Lemma (ESL) files of the CELEX database. We also defined Age of Acquisition features based on the psycholinguistic database compiled by Kuperman et al. (2012). Finally, we included the ratios of various POS tags to the total number of words as POS density features, using the POS tags from the Berkeley parser output.

6.4.6 Other

Lemma realization We specified a set of features that is calculated for each distinct lemma and three feature sets generalizing over all lemmas of the same category (*lm. realiz.*):

1. Distinct lemma counts of a specific category normalized by the total count of this category in a document. For example, if the lemma *can* is found in a document two times as a verb and five times as a noun, and the document contains 30 verbs and 50 nouns, we obtain the two features *can:VB = 2/30* and *can:NN = 5/50*.
2. Type-Lemma ratio: lemmas of same category normalized by total lemma count
3. Type-Token ratio: tokens of same category normalized by total token count
4. Lemma-Token Ratio: lemmas of same category normalized by tokens of same category

Proficiency and prompt features Finally, for some settings in the *closed* task we also included two nominal features to encode the proficiency (low, medium, high) and the prompt (P1–P8), provided as meta-data along with the T11 corpus.

6.5 Evaluation Setup

We developed our approach with a focus on the *closed* task, training the models on the T11 *train* set and testing them on the T11 *dev* set[6]. Based on Bykh et al. (2013), for the *closed* task, we report the accuracies on the T11 *dev* set for all models – single feature and ensemble models as introduced in Section 6.6.1 and Section 6.6.2 –, before presenting the accuracies on the submitted T11 *test* set models, which were trained on the T11 *train* ∪ *dev* set. In addition, for the submitted models we report the accuracies obtained via 10-fold cross-validation on the T11 *train* ∪ *dev* set using the folds specification provided by the organizers of the shared task. Finally, we present the accuracies for single feature models obtained on the T11 *test* set by training the classifier on the T11 *train* ∪ *dev* set – These results were computed after the shared task.

The results for the *open-1* task are obtained by training the models on the NT11 set, and the results for the *open-2* task are obtained by training the models on the T11 *train* ∪ *dev* set ∪ NT11 set. For the *open-1* and *open-2* tasks, we report the basic single feature type results on the T11 *dev* set and two sets of results on the T11 *test* set: the results for the actual *submitted* systems and the results for the *complete* systems, i.e., including the features used in the *closed* task submissions that for the open tasks were only computed after the submission deadline (Given our focus on the *closed* task and finite computational infrastructure).

Below we provide a summary of the various accuracies (%), we report for the different tasks:

- Acc_{test}: Accuracy on the T11 *test* set after training the model on:
  - *Closed*: T11 *train* ∪ *dev* set
  - *Open-1*: NT11 set
  - *Open-2*: T11 *train* ∪ *dev* set ∪ NT11 set
- Acc_{dev}: Accuracy on the T11 *dev* set after training the model on:
  - *Closed*: T11 *train* set
  - *Open-1*: NT11 set
  - *Open-2*: T11 *train* set ∪ NT11 set

[6]The T11 *test* set was not available at this point.

- $Acc^{10}_{train \cup dev}$: Accuracy on the T11 *train* $\cup$ *dev* set obtained via 10-fold cross-validation using the data split information provided by the organizers, applicable only for the *closed* task.

In terms of the tools used for classification, we employed *LIBLINEAR* (Fan et al., 2008) using *L2-regularized Logistic Regression*, *LIBSVM* (Chang & Lin, 2011) using C-SVC with the RBF kernel and *WEKA SMO* (Platt, 1998; Hall et al., 2009) fitting logistic models to SVM outputs utilizing the *-M* option. Which classifier was used where is discussed below.

6.6 Classifier Models

In this section we describe the classifier models used in the given context.

6.6.1 Single Features

We start by evaluating the performance of the different features separately. This was accomplished by training a separate classifier for each feature type. As classifier, we generally used LIBLINEAR, except for *complexity* and *lm. realiz.*, where WEKA SMO performed consistently better. We consider these results the starting point for any further explorations.

6.6.2 Ensembles

After the exploration of the single feature type performance, we turn to combining the different models. First, we are interested in the joint performance of all features used for this study. Second, we investigate the performance of different systems based on various feature combinations to get more insight about the employed feature set.

We followed Tetreault et al. (2012) in exploring two options: On the one hand, we combined the different features directly in a *single vector*. On the other hand, we used an *ensemble* classifier. The ensemble setup we used, combines the probability distributions provided by the individual classifiers for each of the incorporated single feature types. The individual classifiers were trained as described

in Section 6.6.1. The ensembles were trained and tested using LIBSVM, which in our tests performed better for this purpose than LIBLINEAR. To obtain the ensemble *training files*, we performed 10-fold cross-validation for each feature model on the T11 *train* set (for internal evaluation) and on the T11 *train* ∪ *dev* set (for submission) and took the corresponding probability estimate distributions. For the ensemble *test files*, we took the probability estimate distribution yielded by each feature model trained on the T11 *train* set and tested on the T11 *dev* set (for internal evaluation), as well as by each feature model trained on the T11 *train* ∪ *dev* set and tested on the T11 *test* set (for submission).

In our preliminary explorations, the ensemble classifier outperformed the single vector approach, which is in line with the findings of Tetreault et al. (2012). Thus, we focused on ensemble classification for combining the different features.

Then we applied the same system design to the *open-1* and *open-2* tasks using as many feature models as were available in time for the submission. For the *open-1* task we used the NT11 set and for the *open-2* task we used the NT11 set ∪ T11 *train* ∪ *dev* sets to obtain the ensemble *training files* and correspondingly the T11 *test* set to obtain the ensemble *test files*.

6.7 Results

In this section, we provide our results for the individual feature types as well as for the ensemble models, and we describe our systems submitted to the First NLI Shared Task.

6.7.1 Single Features

Table 6.2 presents Acc_{dev} and Acc_{test} single feature type results in the context of the *closed* task as introduced in Section 6.5, thus allowing for direct comparisons between the outcomes on the T11 *dev* vs. T11 *test* sets. Some of the features show slightly better Acc_{dev} results, others yield slightly higher Acc_{test} outcomes. In general, the performance on the *dev* set is comparable to the performance on the *test* set, thus the T11 corpus seems to be sufficiently uniform.

feature type	*closed* task	
	Acc_{dev}	Acc_{test}
1. rc. word ng.	***81.3***	***79.6***
2. rc. OCPOS ng.	67.6	67.7
3. rc. word dep.	67.7	68.0
4. rc. func. dep.	62.4	61.0
5. complexity	37.6	39.5
6. stemsuffix bin.	50.3	50.2
7. stemsuffix cnt.	48.2	49.8
8. suffix bin.	20.4	18.6
9. suffix cnt.	19.0	22.1
10. type dep. lm.	67.3	65.1
11. type dep. POS	46.6	46.8
12. local trees	49.1	49.8
13. dep. num.	39.7	38.8
14. dep. var.	41.5	40.9
15. dep. POS	47.8	47.2
16. lm. realiz.	70.3	69.6

Table 6.2: Single feature type results on T11 *dev* set (Acc_{dev}) and T11 *test* set (Acc_{test}) in the context of the *closed* (main) task, where only T11 data was used.

The Acc_{dev} single feature performance across the tasks is shown in the Table 6.3. The results reveal some interesting insights into the employed feature sets. The figures show that the recurring word-based n-grams (*rc. word ng.*) explored in Chapter 5, constitute the best performing single feature in our set, consistently yielding highest accuracies across the tasks (up to 81.3%). This finding confirms our conclusions in Chapter 5, being well in line with the previous research on different data sets, showing that lexical information seems to be highly relevant for the task of NLI (Brooke & Hirst, 2011; Bykh & Meurers, 2012; Jarvis et al., 2012b; Jarvis & Paquot, 2012; Tetreault et al., 2012). Interestingly, also the more abstract linguistic features in our set – first of all, different features based on dependencies, as well as local trees and lemma realization features –, seem to contribute relevant information, considering the chance baseline of 9.1%.

feature type	Acc_{dev} *closed*	*open-1*	*open-2*
1. rc. word ng.	***81.3***	***42.0***	***80.3***
2. rc. OCPOS ng.	67.6	26.6	64.8
3. rc. word dep.	67.7	30.9	69.4
4. rc. func. dep.	62.4	28.2	61.3
5. complexity	37.6	19.7	36.5
6. stemsuffix bin.	50.3	21.4	48.8
7. stemsuffix cnt.	48.2	19.3	47.1
8. suffix bin.	20.4	9.1	17.5
9. suffix cnt.	19.0	13.0	17.7
10. type dep. lm.	67.3	25.7	67.5
11. type dep. POS	46.6	27.8	27.6
12. local trees	49.1	26.2	25.7
13. dep. num.	39.7	19.6	41.8
14. dep. var.	41.5	18.6	40.1
15. dep. POS	47.8	21.5	47.4
16. lm. realiz.	70.3	30.3	66.9

Table 6.3: Single feature type results on T11 *dev* set (Acc_{dev}), using different training sets according to the definition of the particular tasks.

Having explored the performance of the single feature models, the following three points seem particularly interesting:

1. Is it possible to outperform the recurring word-based n-grams as the best performing feature type, by incorporating the other features from our set?
2. What is the joint performance of our whole feature set?
3. What is the performance of different subsets of our features?

Thus, in the following section, we investigate different feature combinations based on the feature set introduced above.

6.7.2 Ensembles

In this section, we explore combining single feature models in the context of the different tasks, i.e., the systems presented below reflect our submission to the

First NLI Shared Task competition. In particular, we employed the systems listed in Table 6.4, which we chose in order to test all features together (1), the best performing single feature (2), everything except for the best single feature (3), and two subsets, with the former performing well in our preliminary experiments (4) and the latter primarily combining some more abstract linguistic features of diverse nature (5). In the following result tables and explanations, the system ids in the table headers correspond to the ids in Table 6.4, and the symbols have the following meaning:

- x = feature type used
- - = feature type not used
- -* = feature type ready after submission (concerns the *open* tasks)

id	system description	system type
1	overall system	ensemble
2	rc. word ng.	single model
3	#1 minus rc. word ng.	ensemble
4	well performing subset	ensemble
5	"linguistic subset"	ensemble

Table 6.4: Explored systems for all three tasks

***Closed* (main) task (single-corpus)** We submitted the predictions for all systems listed in Table 6.4. The results are presented in Table 6.5.[7]

We report the Acc_{test}, Acc_{dev} and $Acc^{10}_{train \cup dev}$ accuracies as introduced in the Section 6.5. Different from the single feature outcomes presented in Section 6.7.1, where the Acc_{dev} and the Acc_{test} performance was comparable, for ensemble settings the Acc_{dev} results are consistently better than Acc_{test}. This shows that the findings based on single features are not directly transferable to the ensembles. The cross-validation results $Acc^{10}_{train \cup dev}$ are comparable to the Acc_{test} outcomes, and are again consistently worse than the Acc_{dev} outcomes. Thus, the *dev* set seems to exhibit some properties which make it on average easier to classify.[8]

[7]Here and in the following result tables, the best result on the *test* set, which is the most relevant result in the context of the shared task, is shown in bold.

[8]That should be taken into account, if the *dev* set is used for tuning the parameters of a system.

feature type	systems 1	2	3	4	5
1. rc. word ng.	x	x	-	x	-
2. rc. OCPOS ng.	x	-	x	x	-
3. rc. word dep.	x	-	x	x	-
4. rc. func. dep.	x	-	x	x	-
5. complexity	x	-	x	x	x
6. stemsuffix bin.	x	-	x	x	x
7. stemsuffix cnt.	x	-	x	-	x
8. suffix bin.	x	-	x	x	x
9. suffix cnt.	x	-	x	-	x
10. type dep. lm.	x	-	x	-	x
11. type dep. POS	x	-	x	-	x
12. local trees	x	-	x	-	x
13. dep. num.	x	-	x	x	-
14. dep. var.	x	-	x	x	-
15. dep. POS	x	-	x	x	-
16. lm. realiz.	x	-	x	x	-
proficiency	x	-	x	x	-
prompt	x	-	x	x	-
Acc_{test}	**82.2**	79.6	81.0	81.5	74.7
Acc_{dev}	85.4	81.3	83.5	84.9	76.3
$Acc^{10}_{train \cup dev}$	82.4	78.9	80.7	81.7	74.1

Table 6.5: Results for the *closed* task

Overall, comparing the results for the different systems shows the following main points[9]:

- The overall system performed better than any single feature alone (cf. Table 6.2, Table 6.3 and Table 6.5). The ensemble thus is successful in combining the strengths of the different features.

- The *rc. word ng.* feature alone (2) performed very well, but the overall system without that feature type (3) still outperformed it. Thus apparently the different properties accessed by more elaborate linguistic modelling contribute some information not provided by the surface-based n-grams.

[9]The ids of the corresponding systems are provided in parentheses.

- A system incorporating a subset of the different features (4) performed still reasonably well. Hence, it is conceivable that a subsystem consisting of some selected features would perform equally well (eliminating only information present in multiple features) or even outperform the overall system (by removing some noise). We investigate that point in more detail in Part IV.

- Combining a subset of features where each feature incorporates some degree of linguistic abstraction (5) – in contrast to pure surface-based features such as word-based n-grams – performed at a reasonably high level. It supports the assumption that incorporating more linguistic knowledge into the system design can provide a quantitative edge.

Finally, putting our results into the context of the First NLI Shared Task, with our best Acc_{test} value of 82.2% for *closed* as the main task, we ranked fifth out of 29 participating teams. The best result in the competition, obtained by the team *JAR*, is 83.6% (see Section 3.2).[10]

***Open-1* task (cross-corpus)** As for the *closed* task, we explored using the systems listed in Table 6.4. However, different from the *closed* task, here we report two different Acc_{test} values: The accuracy for the actual *submitted* systems (Acc_{test}) and for the corresponding *complete* systems (Acc_{test} with $*$) as explained in Section 6.7.2.

Conceptually, the *open-1* task is a traditional cross-corpus task, where we used NT11 for training and the T11 *test* set for testing. In general, we observe a drop in accuracy by around $\frac{1}{2}$. This setting is more challenging for several reasons. First, the amount of data, we were able to obtain to train our model is far below what was provided for the *closed* task. Thus some drop in accuracy was expected. However, the smaller data size is hardly solely responsible for such a big performance decrease. Tetreault et al. (2012) explored the interplay between the corpus size and the accuracy using an extended version of the T11 corpus, namely the *TOEFL-Big*

[10] According to the significance testing provided by the shared task organizers, the given difference of 1.4% is not statistically significant, i.e., $p = 0.124$ for a pairwise comparison using McNemar's test.

feature type	systems 1	2	3	4	5
1. rc. word ng.	x	x	-	x	-
2. rc. OCPOS ng.	x	-	x	x	-
3. rc. word dep.	x	-	x	x	-
4. rc. func. dep.	x	-	x	x	-
5. complexity	x	-	x	x	x
6. stemsuffix bin.	x	-	x	x	x
7. stemsuffix cnt.	x	-	x	-	x
8. suffix bin.	x	-	x	x	x
9. suffix cnt.	x	-	x	-	x
10. type dep. lm.	-*	-	-*	-	-*
11. type dep. POS	-*	-	-*	-	-*
12. local trees	-*	-	-*	-	-*
13. dep. num.	x	-	x	x	-
14. dep. var.	x	-	x	x	-
15. dep. POS	x	-	x	x	-
16. lm. realiz.	x	-	x	x	-
Acc_{test}	36.4	**38.5**	33.2	37.8	21.2
Acc_{test} with $*$	37.0	n/a	35.4	n/a	29.9

Table 6.6: Results for the *open-1* task

corpus (see Section 3.1), employing a wide range of features. The findings suggest that reducing the T11 data to a size comparable to our NT11 corpus is not likely to cause such a huge drop in accuracy. Second, the uneven distribution of the texts among the different L1s in the NT11 corpus might have a negative effect. Third, the models are trained on data that is likely to differ from the *test* set in a number of parameters, including possible differences in genre (cf. Chapter 5), task and topic (Brooke & Hirst, 2011, 2012b), or proficiency level (Tetreault et al., 2012) – We believe that these issues contribute significantly to the observed decrease in accuracy. Particularly interesting are the following findings:

- Our best accuracy of 38.5% for this task was obtained using the *rc. word ng.* feature alone (2). Thus, adding the more abstract features did not improve the accuracy. The differing properties of the used corpora discussed above seem to prevent a better generalization. Eventually, this finding confirms

our conclusions in Chapter 5, suggesting that lexical features, such as word n-grams, seems to play a key role for the cross-corpus performance in NLI.

- The system combining a subset of features (4) outperformed the overall system (1). This finding supports our assumption, made in the discussion on the *closed* task (see the paragraph above) that the ensemble classifier can be optimized by selective model combination instead of combining all available models.

Finally, putting our results into the context of the First NLI Shared Task, our best Acc_{test} value of 38.5% for the *open-1* task achieved rank two out of three participating teams. The best accuracy of 56.5% was obtained by the team *TOR* (see Section 3.2). While the *open-1* task results, in general, are much lower than the *closed* task results, highlighting an important challenge for future NLI work (see Section 3.3), they nevertheless constitute important steps forward, taking into account the chance baseline of 9.1%.

***Open-2* task (extended single-corpus)** For the *open-2* task we provide the same information as for *open-1*. The results are presented in Table 6.7. Again, we report two different Acc_{test} values: The accuracy for the actual *submitted* systems (Acc_{test}) and for the *complete* systems (Acc_{test} with $*$).

For the *open-2* task, we put the T11 *train* $\cup$ *dev* and NT11 sets together to train our models. The interesting question behind this task is, whether it is possible to improve the accuracy of NLI by adding data from corpora other than the one used for testing. This is far from obvious, especially considering the low results obtained for the *open-1* task pointing to significant differences between the T11 and the NT11 corpora. Essentially, the *open-2* task also is closest to the real-world scenario of using whatever resources are available to obtain the best result possible. Particularly interesting are the following findings:

- Similar to the *closed* task, the overall system (1) performed better than any single feature alone (see Tables 6.3 and 6.7). The ensemble thus is again successful in combining the strengths of the different features.

- When using all features combined (1), our results for the *open-2* task is 84.5%, and thus better than those we obtained for the *closed* task (82.2%). So adding data from a different domain improves the results. This is encouraging, since it indicates that using our feature set, something general about the language used is being learned, not (just) something specific to the T11 corpus.

Finally, putting the results into the context of the First NLI Shared Task, our best Acc_{test} value of 83.5% (84.5% for Acc_{test} with $*$) is the highest accuracy for the *open-2* task, i.e, we obtained the first rank out of four participating teams in this task (see Section 3.2).

	systems				
feature type	1	2	3	4	5
1. rc. word ng.	x	x	-	x	-
2. rc. OCPOS ng.	x	-	x	x	-
3. rc. word dep.	-*	-	-*	-*	-
4. rc. func. dep.	x	-	x	x	-
5. complexity	x	-	x	x	x
6. stemsuffix bin.	x	-	x	x	x
7. stemsuffix cnt.	x	-	x	-	x
8. suffix bin.	x	-	x	x	x
9. suffix cnt.	x	-	x	-	x
10. type dep. lm.	-*	-	-*	-	-*
11. type dep. POS	x	-	x	-	x
12. local trees	x	-	x	-	x
13. dep. num.	x	-	x	x	-
14. dep. var.	x	-	x	x	-
15. dep. POS	x	-	x	x	-
16. lm. realiz.	x	-	x	x	-
Acc_{test}	83.5	81.0	79.3	82.5	64.8
Acc_{test} with $*$	**84.5**	n/a	83.3	82.9	79.8

Table 6.7: Results for the *open-2* task

6.8 Conclusions

We explored the task of NLI using a range of different features in the context of the First NLI Shared Task. We considered pure surface features, namely, recurring word-based n-grams, which we investigated in Chapter 5, as our basis. We then explored the contribution and usefulness of some more elaborate, linguistically-motivated features for the task at hand. Our feature set included some features that are novel for the task of NLI, such as recurring function based dependencies, dependency realization features, or different suffix-based features, as well as a range of features, previously used for NLI. Using an ensemble model combining features based on POS, dependency and constituency trees as well as lemma realization, complexity and suffix information, we were able to outperform the high accuracy achieved by the surface-based recurring n-grams alone. The exploration of more elaborate linguistically-informed features thus is not just of theoretical interest, but can also make a quantitative difference for obtaining state-of-the-art performance.

Furthermore, our findings suggest that it seems possible to optimize ensemble classifiers by selective model combination instead of combining all available information. It could reduce the noise or eliminate potentially redundant information from the feature set, and thus, make the systems more efficient.

Finally, based on our findings, the following procedure seems reasonable in building NLI systems:

1. Start with surface features, such as word-based n-grams.
2. Add some more elaborated linguistically-motivated features capturing potential L1-transfer effects at different linguistic levels.
3. Optimize the system, e.g., using model selection.

Connecting the findings and the research questions Regarding the particular five research questions in the focus of this thesis (see Section 1.3), the findings in this chapter contribute to each of them as follows:

1. [LINGUISTIC-FEATURES]: In this chapter, we implemented and explored a range of features, previously used as well as new for the task of NLI. In par-

ticular, we explored the following features, novel for NLI (see Section 6.4 and Tables 6.2, 6.3):

- *Dependency*:
  - Recurring word-based dependencies (*rc. word dep.*)
  - Recurring function-based dependencies (*rc. func. dep.*)
  - Dependency number (*dep. num.*)
  - Dependency variability (*dep. var.*)
  - Dependency POS (*dep. POS*)
- *Morphology*:
  - Suffixes (*suffix bin./cnt.*)
  - Morphologically complex uni-grams (*stemsuffix bin./cnt.*)
- *Complexity*:
  - Language complexity (*complexity*)
- *Other*:
  - Lemma realization (*lm. realiz.*)

In general, these features show accuracies well beyond the chance baseline[11], and adding them to pure surface-based ones such as word-based n-grams, mostly showed an performance increase. Thus, the findings in this chapter suggest that various features on different levels of linguistic modelling are useful for the task of NLI. Exploring them is not only of theoretical interest, but also provides quantitative advantages by increasing the accuracy of the system.

2. [CROSS-CORPUS]: The findings in this chapter suggest that in general high cross-corpus performance is challenging even if a broad feature set is employed. Our best cross-corpus accuracy was obtained using word-based n-grams (*rc. word ng.*) alone, confirming our findings in Chapter 5 on the role of lexical features for the cross-corpus performance in NLI. Different from

[11]Except for the *suffix bin.* feature type, performing at the chance baseline level in the (cross-corpus) *open-1* task.

the single-corpus findings, combining surface n-grams with more abstract features did not improve the result. However, except for the *suffix bin.* feature, all linguistically-motivated features performed well beyond the chance baseline, which is in line with our single-corpus findings and shows that such features can still be useful in cross-corpus settings. There might be different reasons for the lower performance compared to the single-corpus experiments. First of all, the different parameters varying across the corpora, such as the genre, task and topic, or proficiency, make the task more difficult. Also the smaller data size of the NT11 corpus compared to the T11, and especially the fact that for some L1s (such as Hindi or Telugu) it is hard to obtain a reasonable amount of appropriate training data, certainly plays some role. High cross-corpus performance remains a general challenge in NLI (see Section 3.3.1).

3. [MODEL-OPTIMIZATION]: We conducted first experiments on combining different features. Our findings suggest that it seems possible to optimize ensemble classifiers by selective model combination instead of combining all available information: In some cases, subsets of the features perform better than models combining all of the features (see Section 6.7.2). We focus on this issue in Part IV.

4. [GENERAL-STRUCTURES]: In this chapter, we explored a range of features, incorporating linguistic abstractions. E.g., the *rc. func. dep.* features abstract over particular *rc. word. dep.* features, by using grammatical functions instead of words to represent dependents. In general, such features are capable of reflecting more general linguistic structures and thus are especially interesting from the qualitative point of view. However, higher abstraction, in general seems to come with a lower performance level. We observe that phenomenon also when comparing the *rc. word. ng.* and the *rc. OCPOS ng.* features (cf. Chapter 5). We assume, the reason for this is that for the more abstract features, the classifier lacks the access to some of the very specific, potentially highly indicative surface traits. Thus, the benefits of potentially lower data sparsity seem to get outweighed by the loss of some indicative surface cues. From the quantitative point of view,

it seems generally reasonable to combine the more abstract features with some surface-based ones complementing each other.

5. [VARIATIONIST-PERSPECTIVE]: First of all *dep. var.*, but also *dep. num.* and *dep. POS* are our first features that were inspired by the variationist perspective on the problem (see Section 2.2), i.e., they encode the writer choices in a given frame. All of them show a performance that is well beyond the chance baseline across the settings (see Table 6.2 and Table 6.3). It seems worthwhile to further pursue that research direction which is promising from the quantitative and the qualitative perspectives: It has the potential to improve the NLI performance based on potentially indicative preferences, as well as to provide interpretable qualitative results by unveiling the particular choices made by learners with different L1s. We focus on this research direction in Part III.

Part III

A Variationist Approach to NLI

Chapter 7

Introduction

In this part of the thesis, we discuss how a particular linguistic theory, namely, the *variationist sociolinguistics* perspective (see Section 2.2), can be applied to a NLP task such as NLI, and what are the potential advantages and limitations of the approach. In particular, after clarifying some general issues in Chapter 7, we describe and discuss our implementation of the approach in Chapter 8. Then, we conduct a set of quantitative experiments in Chapter 9, exemplifying potential performance gains. Further, we provide first qualitative findings in Chapter 10 and show how the method can be used to advance the insight in SLA. Finally, we discuss the advantages and limitations of the approach in Chapter 11.

7.1 A Variationist Perspective on NLI

In general, the language offers many different ways for expressing meanings and intentions. Thus, whenever we speak or write, we usually have to make certain choices regarding the particular lexical items and the linguistic structure of the output. In other words, in a given context we have to pick a *particular option* or a *variant* from a set of *possible options* constituting a particular *variable*. If there is a set of possible options, then some speakers might tend to prefer some of them, while avoiding others. Research in VS suggests that the preference for an option may depend on a range of factors such as the social status, gender, age or ethnicity (Labov, 1972; Tagliamonte, 2012; Oliva & Serrano, 2013; Geeslin & Long, 2014).

In addition, following the idea of the VS, recent research in the language learning context argues that a preference for particular options can be indicative of individual characteristics such as the language proficiency or L1 (Callies & Szczesniak, 2008; Callies & Zaytseva, 2011; Lüdeling, 2011; Meurers et al., 2014). Since ethnicity and the L1 as a substantial part of it (Coulmas, 1997) are factors that seem to influence the preference for particular variants in the language productions, it seems worthwhile to further explore that research direction in the context of NLI, providing a suitable test bed. Applying a variationist perspective to NLI might show advantages in both, quantitative and qualitative regards:

1. It could provide quantitative advantages by facilitating the identification of indicative choices made by writers with different L1s, and then exploiting this information to improve the classification models.

2. It could foster the qualitative analysis by unveiling particular indicative choices that can be analysed within the linguistic theory.

In order to be able to apply the variationist perspective, we have to clarify some core questions, which we list in the next section.

7.2 Implementing a Variationist Perspective: Core Questions

In this part, we explore applying the variationist perspective to the task of NLI using a range of linguistic features, we consider suitable for the given problem. In this regard, the following core questions have to be clarified:

1. What is a suitable definition of a *linguistic variable*?

2. What different *types of linguistic variables* should be distinguished in the given context?

3. How can we *abstract over individual linguistic variables* to obtain insights into more general underlying linguistic structures reflected in NLI?

4. What *linguistic variables to explore* as features in the context of this thesis?

Essentially, giving answers to these questions constitutes subsequent steps in the feature engineering process (see Section 2.3). At the end of this process, we will implement and evaluate systems based on different features, which we call *variationist features*, and discuss their use for NLI.

Finally, in referring to the provocative question in the title of this thesis, we try to approach the general question of how much is the choice of particular variant of a linguistic variable dependent on the actual L1 of the individual in its non-native productions? Are the choices we make, maybe, strongly predetermined by our L1, so that in practice there is essentially no real choice? A comprehensive answer to that question is clearly beyond the scope of a single thesis. Nevertheless, the results of our study contribute a piece to its clarification.

7.3 Relevant Related Work in NLI

Related work overview To the best of our knowledge, the first work approaching the particular task of NLI by explicitly considering language variation was Krivanek (2012). The author investigates syntactic alternations as characteristic features of learner language. In particular, she explores using theory- as well as data-driven verb alternations as features for NLP tasks such as NLI. In Meurers, Krivanek & Bykh (2014) we partially refer to the findings in Krivanek (2012) and consider them in a broader context, stressing the importance of linguistic abstraction for tasks such as NLI. In Bykh & Meurers (2014) we systematically explore variationist feature encodings using non-lexicalized and lexicalized CFG production rules as features for NLI. We show that such features can provide a quantitative edge. Following Meurers, Krivanek & Bykh (2014), in Bykh & Meurers (2016) we further explore verb subcategorization features under a varitionist perspective, and propose a way of abstracting from individual features to classes using feature grouping. We show that this can optimize the classification models and enhance the qualitative analysis. In sum, the findings show that a variationist approach to NLI can indeed provide quantitative and qualitative advantages.

Krivanek (2012) as basis for our variationist approach In the following we briefly describe the core issues discussed in Krivanek (2012), which provides a starting point for the variationist approach proposed and explored in this thesis.

In this study, the author focuses on exploring theory- and data-driven verb alternations as features for NLI. The theory-driven alternations are implemented based on the theory of English verb classes and alternations, proposed by Levin (1993). The author distinguishes between alternations such as Preposition Drop Alternations (e.g., *Jim met with Christian.* vs. *Jim met Christian.*), Dative Alternation (e.g., *Bill sold a car to Tom.* vs. *Bill sold Tom a car.*), Locative Alternations (*Jack sprayed paint on the wall.* vs. *Jack sprayed the wall with paint.*), Creation and Transformation Alternations (*Martha carved a toy from the piece of wood.* vs. *Martha carved the piece of wood into a toy.*, and *He turned into a frog.* vs. *He turned from a prince into a frog.*), etc. Krivanek (2012) follows the taxonomy and terminology suggested by Levin (1993). At the same time, she points out that a strict implementation of Levin's theory is hardly feasible in the context of an automatic approach, first of all, because of the *meaning equivalence* assumption regarding the different syntactic constructions within an alternation. There is only a limited availability of syntactically different utterances with the same meaning in the used corpora, i.e., it is hard to discover them, because of the lack of appropriate semantic annotation. Thus, Krivanek (2012) suggests to restrict the concept of alternation to the verb's valence variation.

The author suggests three types of features based on verb alternations:

1. The syntactic patterns as separate features (*Pattern*)

2. Each of the syntactic patterns combined with each of the particular verbs it occurs with (*Verb+Pattern*)

3. Each of the syntactic patterns combined with a group of verbs constituting the same alternation (*Alternation*)

The features were encoded using relative frequencies of the patterns. For the feature type (2) the values are calculated relative to the particular verb lemmas, the different patterns are occurring with. For the feature type (3), where we have

to calculate the relative frequencies of patterns over a set of verbs, *micro-* and *macro-average* calculation was employed:

- *Micro-Average*: Relative frequency of a pattern p, calculated by summing up the frequency of p across all verb lemmas belonging to a particular alternation A, and normalizing that value by the overall frequency of the alternation A (i.e., the frequency of all patterns across all verb lemmas in A).
- *Macro-Average*: Relative frequency of a pattern p, calculated by summing up the relative frequencies of p for each verb lemma belonging to the alternation A separately, and normalizing that value by the number of verb lemmas in A.

The features were generated based on CFG parses using ICLE and LOCNESS corpora. For a binary classification task, namely L1 Chinese learners vs. native English speakers, the system showed an accuracy of 53.3% for the feature type (1), 73.3% for the feature type (2), as well as 63.3% employing micro-average and 59.2% employing macro-average for the feature type (3). The author also provides some examples for alternations where there seem to be different preferences by writes with a different L1 regarding the possible options within an alternation. For example, it turned out that in the context of the "Locative preposition drop alternation" (e.g., *Martha climbed up the mountain* vs. *Martha climbed the mountain*), L1 Chinese learners tend to prefer the variant without the locative preposition compared to the native English speakers.

Next, the author proposes a data-driven approach based on automatic alternations. Here the alternations are compiled based on the actual data, i.e., based on all verb lemmas and all syntactic patterns that occur with the individual verb lemmas in the given data. The author refers to the data-driven syntactic patterns as *subcategorization (subcat) patterns*. Since some verb lemmas can occur with many different subcat patterns, Krivanek (2012) suggests to option for compiling alternations:

- *Single-Verb Alternation*: Each verb lemma belongs to a single automatic alternation, consisting of verb lemmas which occur with exactly the same subcat patterns in the data.

- *Duplicate-Verb Alternation*: Each verb lemma can belong to different automatic alternations, depending on which of the possible subcat pattern subsets (drawn from the whole subcat pattern set for a particular verb lemma) is considered.

The data-driven approach outperformed the theory-driven approach across the different feature types: For the feature type (1) the accuracy was 81.7% and for the feature type (3) the best accuracy was 83.3% using macro-average and duplicated-verb alternations. The result for the feature type (2) is not reported for the binary data set, but the author reports corresponding results for a multi-language setting with five different L1s: The accuracy for the feature type (2) is 75.3%, which is the best outcome in this setting – The feature type (1) performed at 52% and the feature type (3) at 62.7%. The author also shows that using bigger texts, obtained by merging multiple original texts together, further improves the accuracy by reducing data sparsity issues.

Krivanek (2012) concludes that both, the theory- and data-driven alternations have good reason to be used as features for NLI, but for different reasons: The data-driven features yield high accuracies, but at the same time they are rather difficult to interpret qualitatively. Whereas the theory-driven features are performing at a lower level, but at the same time they are easier to put in the context of the linguistic theory and thus, seem to be better suitable for qualitative analysis.

Discussion The approach by Krivanek (2012) is a solid work connecting a language variation perspective using verb alternations as features and the task of NLI. It provides a good starting point for further developments and investigations. Based on that work, in this chapter we develop a general variationist approach that can be applied to NLI and any other NLP task potentially benefiting from the variationist gist. In the following, we discuss the main points related to Krivanek's study which from our point of view could be improved or extended, and which we will thus target in the course this thesis.

The work by Krivanek (2012) constitutes a study in the context of language variation, but it is not explicitly considered as a contribution to the field of VS. Thus, the contribution does not discuss some of the issues relevant in this context,

such as the general notion of the linguistic variable suitable for the application in this and other related tasks, the different variable types reasonable to distinguish in the given context, etc. Further, it only considers a specific sort of verb alternation features using a relatively small data set. The approach also lacks a general, flexible technique for grouping features, and thus exploring different levels of abstraction, which can be beneficial in quantitative (less data sparsity) as well as qualitative (more general insights) regards. Finally, the study does not provide an analysis of the data-driven outcomes from the qualitative point of view. However, exploring such features in detail has the potential of revealing interesting new patterns in the language use of individuals with differing characteristics such as the L1 of the writers.

We already established first relevant links between Krivanek (2012) and the VS perspective in Meurers, Krivanek & Bykh (2014), and our investigations in this part of the thesis further extend this perspective in a way supporting answers to the core questions we formulated in Section 7.2.

Chapter 8

Variationist Feature Engineering

8.1 Linguistic Variables Revisited

The first core issue to clarify is the question of what definition of a linguistic variable is suitable and applicable in the context of this thesis? Here we face a conceptual issue. The traditional, strict definition of a linguistic variable is "two or more ways of saying the same thing" (Section 2.2). Thus, it is based on the *meaning*, assuming *semantic equivalence* of the variants. At the level of phonetics-phonology, it seems possible to comply with it – Even if the words are pronounced differently by different speakers, their meaning is still the same. However, in practice the requirement of semantic or functional equivalence is rather difficult to sustain across different linguistic levels (Tagliamonte, 2012; Oliva & Serrano, 2013; Krivanek, 2012):

> "Establishing functional equivalence beyond the level of phonetics-phonology is problematic. Lay people and linguists alike will argue strongly for meaning differences when presented with potential variables, even when they are framed in near identical phrases. Do the [following] two sentences mean the same thing?
>
> (a) I think she*'ll be* cheeky. [...]
>
> (b) I think she's *gonna be* pretty cheeky. [...]"
>
> (Tagliamonte, 2012, p. 16)

In the example above, the utterances are clearly *similar* in their meaning. However, suggesting that there is an absolute *equivalence* would be a hardly sustainable claim – The two utterances are not freely interchangeable in any possible context. In fact, semantic equivalence beyond the level of phonetics-phonology is rather an alleged equivalence, because there still can be perceived differences in the social or geographic distribution, the association to particular contexts of human interaction, the degree of newness or the genre, etc. (Oliva & Serrano, 2013; Weber, 2012). Moreover, some recent research in VS suggests that strictly requiring semantic equivalence of the variants as a prerequisite might be not the best option in any context:

> "Grammatical variants may not be equivalent in meaning [...]. But, in our view, the important question would be: In what sense does this make it impossible to analyze syntactic variability?
>
> [...] The researchers who first extended the use of variationist tools to the study of syntax were not wrong in doing so; their only mistake was probably to assume the philosophy together with the method, accepting synonymy as a prerequisite whose absence would preclude an approach to syntactic usage as variation proper. Meaning differences are not the problem but the solution: they are indeed what justifies the analysis of syntactic variation as a useful task towards the goal of achieving better knowledge of human communication.
>
> [...] Meaning differences need not to be an obstacle for research if it is accepted that linguistic varieties, whether of a geographical, social or any other kind, differ not just in their tendency to choose particular forms, but also in their preference for the kinds of meanings conveyed by such forms."
>
> (Oliva & Serrano, 2013, pp. 19, 20-21 and 63)

Thus, exploring meaning differences and preferences under a variationist perspective seems to be a interesting research area, mostly ignored in the variationist context so far.

Besides the conceptual issues, restricting the feature space to semantically equivalent cases only, would pose a substantial limitation for a data-driven approach, which apart from a interest in the qualitative insight, is also aiming at the exploration of the potential quantitative advantages of applying a variationist perspective to the task of NLI. Even if in some cases we decide to argue for semantic equivalence, it seems to be a rare phenomenon though (cf. Krivanek, 2012). Since the data for the task at hand is rather limited, it would lead to serious data sparsity issues, usually resulting in low performance. We conclude that if our interest goes beyond the phonetic-phonology level, keeping the traditional definition of the linguistic variable is in practice hardly sustainable for NLP tasks such as NLI.

Thus, in order to implement a variationist perspective on the given problem though, we have to revise the basic definition of the linguistic variable and adapt it to a version that on the one hand, is suitable for the given problem, and at the other hand, is acceptable from the theoretical point of view. Here, the notion of *weak complementarity*, introduced by Sankoff & Thibault (1981), becomes an attractive starting point. It is based on the observation that there is a relationship between some options, in that where one variant is used more frequently, the other one is less frequent. These variants do not necessarily have the same meaning. The idea of weak complementarity can be summarized as follows:

> "This is the idea that linguistic variables can be identified by their distribution across the speech community rather than by the fact that they mean the same thing. [...] In reality, an LVC [Language Variation and Change] analysis begins with the observation that where one variant is used more often another variant is used less. When this observation is made of syntactic, semantic or discourse-pragmatic features, form/function correspondence cannot be sustained because variants involved in the same change may not mean precisely the same thing. However, if they are members of the same structured set in the grammar of the speech community these patterns can be *observed*. The criterion for identifying weak complementarity is a correlation between occurrence rates and some extralinguistic factor of individual speakers such as age, sex, or social index."
>
> (Tagliamonte, 2012, p. 16)

Thus, the semantic equivalence component is still present, but it becomes less prominent, i.e., it is not a necessary prerequisite any more. Inspired by the idea of weak complementarity (Sankoff & Thibault, 1981) and based on the assumption that potentially differing meanings of the variants are not an obstacle per se but rather a chance for discovering interesting new patterns in communication (Oliva & Serrano, 2013), we propose the following, more general definition of a linguistic variable:

- *Linguistic Variable*, basic: Two or more ways of saying the same thing (Section 2.2).
- *Linguistic Variable*, revised: Two or more conceptually, structurally or contextually related linguistic variants, showing differing occurrence rates with respect to some extralinguistic factors such as the L1.

While the proposed definition of the linguistic variable does not comply to the variationist perspective in the *most strict*, traditional sense, it is still in line with the general variationist idea in that it considers related linguistic choices made by speakers with respect to a particular extralinguistic factor such as the L1.[1] We follow this more general definition in our implementation of the approach.

In this section, we discussed the notion of a linguistic variable and related issues. Finally, we proposed a revised version of this central notion, which we consider most suitable in the given context. In the following section, we discuss what different types of linguistic variables, are useful to distinguish in this study.

8.2 Types of Linguistic Variables

After having discussed the notion of a linguistic variable in the previous section, in this section, we turn to some details, and discuss a taxonomy of linguistic variables, we consider useful in the context of this thesis.

[1] In fact, the definition incorporates all relevant traditional properties of a linguistic variable listed in Section 2.2, except for the meaning equivalence, discussed in this section.

8.2.1 Relative vs. Absolute

The first distinction to make is between the so called *relative* and *absolute* variables. It reflects the point of how precisely the so-called *principle of accountability*, which is at the heart of the variationist sociolinguistic analysis, is implemented in practice. This central principle can be described as follows:

> "A foundational concept in the Variationist Sociolinguistic approach and one that sets it apart from other methods is the 'principle of accountability' [...]. This is where the analysis begins. Say the analyst is interested in the use of the relative pronoun *who*. The principle of accountability dictates that in addition to examining *who* itself, the analyst must also take into account all the other potential variants within the relative pronoun system. Accountability requires that all the relevant forms in the subsystem of grammar that you have targeted for investigation, not simply the variant of interest, are included in the analysis."
>
> (Tagliamonte, 2012, p. 10)

> "[...] The frequencies achieved by a given form only make sense when put in relation to those of its alleged alternatives."
>
> (Oliva & Serrano, 2013, p. 61)

Relative Variables The *relative variables* strictly implement the principle of accountability, and thus, directly reflect the core idea of the variationist perspective: making a particular choice out of a set of possible options in the given context. It is implemented by calculating the *relative frequency* (i.e., the proportion) of a particular variant with respect to all relevant variants, i.e., all variants constituting a given variable.

Absolute Variables Using relative variables, which best reflects the variationist gist, seems to be the most preferable option. However, it presupposes that a variable can be defined as a closed set of variants which, in fact, is not always possible in practice. Assume, we are interested in the investigation of discourse markers such as *and stuff like that*[2] (Dines, 1980). What exactly is included in the whole set of variants? In such cases it is simply not fully clear. In this regard Oliva & Serrano (2013) states the following:

> "When particular forms are studied whose alternatives are not easy to elucidate, it is always possible to calculate their overall frequencies according to the total word number of the texts under analysis"
>
> (Oliva & Serrano, 2013, p. 65)

Thus, in such cases, it is suggested to calculate simply *normalized frequencies*. The corresponding variables are called *absolute variables* (Oliva & Serrano, 2013, p. 64-67). That is obviously an option that does not comply to the principle of accountability in the strict sense. However, it can be still considered a valid option in the variationist context. It allows to view variation as a creative choice:

> "In our view, absolute variables have crucial implications for a theory of variation. They represent an appropriate methodological concretion of a model viewing variation as a creative *choice* and variants as inseparable form-meaning amalgams with the capacity to communicate something by themselves, not just through their opposition to a number of alternatives"
>
> (Oliva & Serrano, 2013, p. 65, cf. also Coupland (2007))

8.2.2 Lexical vs. Grammatical

Some variables, first of all at the pragmatics level, are related to a concept or a particular discourse function, e.g., discourse markers. Other variables can be described in terms of *lexical forms* or abstract *grammatical categories* (cf. Section 2.2) – a distinction which turns out to be useful in the context of this thesis.

[2] *and stuff like this*, *and stuff*, etc.

Lexical variable This notion depicts a variable where the variants can be related to a particular lexical unit. For example, the synonyms of a word are related to that particular word, thus the corresponding variable can be described by a lexical unit. Another example could be considering different verb lemmas as variables and the various subcategorization patterns, realized by those verb lemmas, as variants (see Section 9.3, cf. Meurers, Krivanek & Bykh, 2014).

Grammatical variable This notion depicts a variable where the variants can be related to a particular grammatical category. For example, we could define a particular POS tag such as pronoun as variable and consider the different surface realizations of that tag in various contexts as variants (see Section 9.2). Another example could be considering a syntactic function such as the subject of a sentence as a variable and the different surface realizations as variants. We could also combine both approaches by restricting the subjects under investigation to a particular POS tag and investigating the corresponding surface realizations (see Section 10.3.1, also cf. Oliva & Serrano, 2013).

8.2.3 Level of Granularity

The distinction described in this section applies in the first place to *lexical variables* as described in Section 8.2.2 which can be realized by variants taken from the same set. We can consider abstracting from the particular lexical variables (e.g., particular verb lemmas which can be realized by various subcategorization patterns as variants) to more general units, i.e., *variable groups*, based on linguistic similarity of the variables (e.g., groups of verb lemmas which can be realized by particular subcategorization patterns as variants). Depending on the abstraction level, we can distinguish different levels of granularity as exemplified below.

Fine-grained Level Variable (FGV) The most fine-grained level is given, if we consider all individual lexical variables separately, i.e., as separate variables. For example, in Krivanek (2012) the *Verb+Pattern* features for a given verb v_i can be viewed as a fine-grained variable, because it is related to a particular single verb (see Section 7.3).

Coarse-grained Level Variable (CGV) The most coarse-grained level is given, if we combine all of the lexical variables in a single abstract variable. For example, in Krivanek (2012) all of the subcategorization patterns together, which constitute the *Pattern* feature, can be viewed as a coarse-grained variable, because it fully abstracts over the particular verbs the patterns are used with (see Section 7.3).

Intermediate-grained Level Variable (IGV) There can be various intermediate levels of abstraction between the two extremes described above. For Example, in Krivanek (2012) the *Alternation* features can be viewed as possible intermediate-grained variables, because they are based on particular sets of verbs, thus being neither fine-grained (the variables are not just individual verbs) nor coarse-grained (there is not only a single abstract variable subsuming all the others) but at some point in between (see Section 7.3). It might well be the case that some sets of variables constituting certain IGVs show a different usage pattern compared to the corresponding CGV, which is maximally abstracting over individual variables. Investigating such differences might provide valuable qualitative insight (see Chapter 10). We propose a flexible approach for generating IGVs at different levels of abstraction in Section 8.3.

8.3 Label-informed Feature Grouping

8.3.1 Introduction

General considerations Exploring usage patterns for certain individual variables, e.g., potential differences in the argument structure realization of particular verbs, can be certainly of interest. However, on the one hand, some individual variables suffer from data sparsity; on the other hand, investigating units beyond individual variables seems to be another attractive option, potentially capable of providing further gains. In sum, abstracting over individual variables might show both, quantitative and qualitative advantages:

- *Quantitative advantages*: Abstracting from individual variables to classes might have a positive effect on the performance:

  - It could reduce potential issues related to data sparsity.
  - It could optimize the models by reducing the feature space.

- *Qualitative advantages*: Abstracting to more general structures is generally valuable for the qualitative analysis, aiming at inferring general patterns and regularities.

In Section 8.2.3, we discussed different levels of abstraction in connection with linguistic variables. The FGVs represent the lowest level of abstraction (i.e., no abstraction), and simply constitute individual linguistic variables, which can be directly accessed. The CGVs represent the highest level of abstraction, and can be obtained by merging all of the suitable linguistic variables into a single set and considering this set as a new individual variable. However, the answer to the question how to obtain IGV variables, i.e., variables at different intermediate abstraction levels, is not obvious. It requires some systematic approach and a concrete technique implementing it. In Bykh & Meurers (2016), we proposed a suitable method, which we describe in very detail in this section.

The gist of the approach Following the general VS idea, we can record the relative frequencies for the different variants used to realize a particular variable in our training data (see Section 8.2.1). Yet, some variables, first of all some individual *lexical variables* (see Section 8.2.2) might occur quite rarely making any generalizations difficult or simply not reliable. At the same time the underlying linguistic structure reflected in some variant preferences might be common to many of such low-frequency variables, showing some general pattern. Even if some individual variables show relatively high frequencies in the data, thus providing a reliable empirical evidence for particular usage patterns, making generalization based on multiple variables showing the same or a similar usage pattern, is certainly still of high interest. In order to discover such more general patterns, inferring *groupings* of variables based on their similarity seems to be a natural solution. For this, first of all, we have to specify what exactly does *similar* mean. Here, we suggest to group together all those variables that show similar variant realization patterns in terms of the *actual variants* and their *frequency distribution*.

However, if variant preferences differ between L1s, which is assumed here, then it is important to keep the distinction between the different L1s in the grouping procedure. Otherwise potentially L1-distinctive information, manifested in the variant preferences, might get lost in a group, if it is build only considering *general* (L1-independent) frequency proportions of the variants in the data. Thus, another important question is how to take the L1 into account when grouping the variables? In other words, how can we group preferably those variables together that for the individual L1s show similar variants realization patterns? For that we have to consider the data for the different L1s *separately* and try to find variables suitable for building groups under the given constraint.

In sum, the gist of the approach can be described as follows:

> *Group together those variables that show the same or a similar variant realization pattern in terms of the actual variants and their frequency distribution, with respect to the individual L1s.*

In the following sections, we propose a method, suitable for implementing the described grouping approach. Since the grouping technique is informed by the L1 labels, we refer to it as *label-informed feature grouping*.

8.3.2 Hierarchical Clustering

The most common method for grouping items[3] is *clustering*, a notion subsuming a class of unsupervised ML techniques (Witten et al., 2011; Alpaydin, 2004, also see Section 2.3). One of the well-established clustering approaches for building feature groups is *hierarchical clustering* (Park, 2013; Krier et al., 2007; Butterworth et al., 2005). In the following we describe the gist and the core components of the hierarchical clustering method based on Alpaydin (2004) and Witten et al. (2011).

Agglomerative vs. divisive clustering In hierarchical clustering, the idea is to build groups of items in a hierarchical manner based on a distance measure. One

[3]For simplicity reasons, we use the general term *item* in this context. In fact, here the items provided to the grouping algorithm are individual variables represented by *feature vectors* encoding the frequency distributions of the corresponding variants.

way to do this is starting with clusters containing a single item, and subsequently building bigger clusters until all items are merged into a single cluster. This describes the so-called *agglomerative hierarchical clustering*, which works bottom-up. There is also a corresponding top-down procedure called *divisive hierarchical clustering*, starting with a single cluster containing all items, and subsequently dividing it in smaller clusters until each of them end up containing a single item.

Distance measures The distance between individual items is usually determined by a standard distance measure such as the *Euclidean* (Eq. 8.1) or *Manhattan* (Eq. 8.2) distance:

$$d_e(x^a, x^b) = \sqrt{\sum_{i=1}^{n} (x_i^a - x_i^b)^2} \tag{8.1}$$

$$d_m(x^a, x^b) = \sum_{i=1}^{n} |x_i^a - x_i^b| \tag{8.2}$$

Linkage methods However, distance measures such as the Euclidean distance only allow for measuring the distance between two single items x^a and x^b, but not between clusters consisting of multiple items. For determining the distance between clusters, so-called *link* or *linkage* methods such as *single-linkage* (Eq. 8.3) or *complete-linkage* (Eq. 8.4) are commonly used:

$$d_{sl}(C_i, C_j) = \min_{x^a \in C_i, x^b \in C_j} d(x^a, x^b) \tag{8.3}$$

$$d_{cl}(C_i, C_j) = \max_{x^a \in C_i, x^b \in C_j} d(x^a, x^b) \tag{8.4}$$

In single-linkage clustering, the distance between two clusters C_i and C_j is defined as the smallest distance between all possible pairs of items $x^a \in C_i$ and $x^b \in C_j$ (the distance between the two closest members of the two clusters).

Whereas in complete-linkage clustering exactly the opposite holds, i.e., the distance between two clusters is defined as the largest distance between all possible pairs of items of the two clusters (the distance between those two items that are most different). Another option is, e.g., *average-linkage* clustering, where the distance between two clusters is defined as the average of the distances between all possible pairs of items of the two clusters, etc.

Dendrogram The output of a hierarchical clustering algorithm is a *dendrogram*, i.e., a binary tree representing the hierarchical structure that reflects the whole clustering procedure. In such a dendrogram, the *leaves* are the individual items that are placed in a sequence on the x axis, and the *branches* show the individual clusters by connecting different elements, which can be realized as follows:

- two individual items[4]
- two clusters of items
- an individual item and a cluster of items

The connections are made at some point, i.e., at some hight on the y axis. The higher the connection point on the y axis, the larger the difference between the elements grouped at this point, and vice versa.

8.3.3 Clustering Linguistic Variables

Core clustering setting We group linguistic variables using the following core setting, which is essentially specified by one of the standard algorithms as well as some widely used standard parameters in hierarchical clustering (see Section 8.3.2):

- Grouping algorithm: *agglomerative hierarchical clustering*

  It was shown in previous research that agglomerative hierarchical clustering can be an effective method for capturing linguistic generalizations. For example, Pate & Meurers (2007) show in the context of

[4]More precisely, two clusters, each consisting of a single item.

PCFG parsing that contextually enriching categories followed by clustering the categories with similar distributions results in a performance improvement. Thus, we opt for utilizing this clustering algorithm as grouping method for our technique.

- Distance measure (parameter): *Euclidean distance*

  We also explored using other distance measures, such as the *Manhattan* or the *Hellinger distance*. The latter is essentially an adaption of the Euclidean distance to probability distributions as input, and thus is generally supposed to be more suitable for measuring the distance between relative variables. However, the Euclidean distance performed best in our preliminary explorations, so we use it for our experiments.

- Linkage method (parameter): *Complete-linkage*

  We also tested other linkage methods such as *single-linkage* or *average-linkage*. However, the outcomes were very similar. Moreover, *complete-linkage*, which is based on the maximum distance between the elements in the clusters to merge (see Section 8.3.2), shows two interesting properties: first, all of the items in the two clusters merged into a new cluster are naturally relatively similar, which is not necessarily the case, e.g., with single-linkage; second, the clusters tend to be compact with small diameters (Witten et al., 2011). These properties make the interpretation of the clusters easier – In general, we can assume higher levels of homogeneity within the clusters. Thus, we kept complete-linkage as parameter.

- Number of clusters c (parameter): c=1

  It means that after clustering we obtain a full *dendrogram*, ending with a single root node, i.e., *a single-rooted binary tree*.

Core procedure Following our considerations in Section 8.3.1, for clustering variables, we can generate a vector representing each variable by the frequency distribution of the corresponding variants in the training data. These vectors can

then be provided as items to the clustering algorithm which outputs a dendrogram showing the similarity between the variables as described in Section 8.3.2.

Adding the label-informed component We can already use the clustering procedure described above to group variables. Now, we still have to clarify how to do the grouping in a way which accounts for the variant preferences in connection with the *different L1s* (see Section 8.3.1). To implement this idea, we do not generate a single vector of variants for a variable, but k vectors, where k is the number of L1 labels in the training data. Each of the k vectors contains the proportions of the variants for a given variable, calculated using the subset of the training data for a particular L1. Then the k vectors for each variable are concatenated in order to get an item (instance vector) for clustering. Hierarchical clustering then groups variables together that for writers of a specific L1 realize a similar set of variants in similar proportions according to our specification in Section 8.3.1. In this way, the feature set for clustering becomes *informed by the L1 labels*, which is the reason why we refer to this technique as *label-informed feature grouping*. In sum, the method follows the variationist perspective and is designed to group those variables together that in terms of their variants behave alike with respect to the classification label.

Let us spell this out in an example using verb lemmas as variables and the corresponding subcategorization patterns as variants. Assume that writers with L1 Spanish prefer the variant $p \in \{p, q\}$, whereas writers with L1 Chinese prefer the variant q in connection with a particular set of verb lemmas V. That information is captured by the difference in relative frequencies for the variants p and q in connection with V in the training data subsets for the two different L1s. Using separate vectors for the different L1s, explicitly provides that relevant information to the clustering algorithm. Clustering thus can identify the group V of verb lemmas that is indicative for the classification purposes in terms of the choice of variants made by different L1s, and that is also of interest from the perspective of interpreting these effects in terms of SLA research.

Table 8.1 further exemplifies the technique using concrete values. Assume some three variables, namely, V_1, V_2 and V_3 occurring at the same frequency $f = 10$ in some training data. Each of them can be realized by one of the two

	label-informed				*plain*	
	$L1_a$		$L1_b$		cross L1	
variable	variant p	variant q	variant p	variant q	variant p	variant q
V_1	4 (0.8)	1 (0.2)	1 (0.2)	4 (0.8)	5 (0.5)	5 (0.5)
V_2	1 (0.2)	4 (0.8)	4 (0.8)	1 (0.2)	5 (0.5)	5 (0.5)
V_3	2 (0.4)	3 (0.6)	5 (1.0)	0 (0)	7 (0.7)	3 (0.3)

Table 8.1: Exemplification of the label-informed clustering advantage. The values are frequencies and corresponding relative frequencies (in parentheses).

possible variants $\{p, q\}$. The values in the table show the frequency distribution for each variable, i.e., the raw frequency and the corresponding relative frequency (in parentheses) for both, the *label-informed* (split by the different L1s, namely, $L1_a$ and $L1_b$) and *plain* (cross L1) settings. The difference in the grouping between the two settings would be as follows:

- *Plain*: The variables V_1 and V_2 would form a group, because the cross L1 frequency distribution for the two variables is exactly the same.
- *Label-informed*: The variables V_2 and V_3 would form a group, because the frequency distribution between the two variables is most similar if considered separately for $L1_a$ and $L1_b$.

Assume that the label-informed procedure formed a cluster consisting of V_2 and V_3 as motivated above. Further assume that the frequency distribution for this cluster in a unseen text t is $p = 0.3$ and $q = 0.7$. This distribution in t is very similar to the distribution for $L1_a$ in connection with V_2 ($p = 0.2, q = 0.8$) and V_3 ($p = 0.4, q = 0.6$) based on the training data provided in the table. At the same time, this distribution in t is substantially different compared to the distribution of the two variables in connection with $L1_b$ ($V_2 : p = 0.8, q = 0.2$ and $V_3 : p = 1.0, q = 0$) in the training data. Thus, a feature based on this group would indicate the label $L1_a$ for t, which is well in line with the general expectations.

Identifying best groups Now, the question is how to decide which groupings are most appropriate? Essentially, it reduces to the question, where to cut the dendrogram in order to obtain reasonable groups of variables? We approach that

issue experimentally, by systematically applying different branch length cut-offs with some step s to the dendrogram, and then evaluating every grouping via text classification, employing the different groupings as features. In general, the step can be selected experimentally or based on the shape of the dendrogram (e.g., depending on the hight of the dendrogram). Here, we opt for $s = 0.1$ as default.

Encoding groups of variables as features In connection with encoding groups of variables as features, there are two questions to clarify. First, how to merge clustered variables with different sets of variants?[5] We suggest the following two options:

- *Union*: One option is to take the union of the variant sets as the resulting variant set for the given variable group. This is our default option for *quantitative* explorations.

- *Intersection*: An alternative is using the intersection of the variant sets. If the intersection for a group is $\emptyset$, then this particular grouping is dropped, i.e., the variables are not merged together. Because some of the variants get dropped here, this option generally leads to higher data sparsity, potentially harming the performance. However, it makes the groups easier to interpret – Eventually, only those variants which occur with all of the variables in a group are used as the variant set for this group. Thus, we prefer this option for the *qualitative* analysis.

Second, how to compute the feature values for such groups? For that we follow Krivanek (2012) and Meurers, Krivanek & Bykh (2014) and use the *micro-average* measure adapted to the variationist perspective (see Section 7.3). We decided against the usage of the *macro-average* encoding, which was also proposed by Krivanek (2012), because it is more sensitive to the occurrence of particular variables in the texts – The normalization in the formula is performed by the number of variables within a group, regardless of which variables from the group are actually occurring in a text.

[5]Clusters of variables showing (slightly) differing sets of variants are well possible with the given clustering procedure.

Overall procedure (putting everything together) In sum, our label-informed feature grouping technique is defined by the following procedure:

1. For each of the n variables V_i and each of the m variants v_j occurring in the whole training data, calculate the matrices M_{ij}^k using the corresponding label-distinct data subset l_k:

$$M_{ij}^k = \frac{f(v_j, V_i, l_k)}{\sum_{q=1}^{m} f(v_q, V_i, l_k)}$$

   where $f(v, V, D)$ yields the frequency of the variant v realizing the variable V in the data D. Here $D = l_k$. If a variant v does not occur in the context of V using data D, $f(v, V, D) = 0$.

2. Perform a horizontal matrix concatenation of the k matrices M_{ij}^k resulting in a single matrix M_{ij} containing the variable based instances for clustering (the rows of M_{ij}).

3. Perform hierarchical clustering with the number of clusters $c = 1$, Euclidean distance, and complete-linkage as parameters.

4. Systematically apply different branch length cut-offs r using a suitable step s. Here we use $s = 0.1$. Evaluate the resulting clusters, i.e., variable groupings by using them as features in an NLI classification setup.

   (a) Merging variables into groups: Let a group (cluster) C contain x variables, each realized by a particular variant set X_i. Then the resulting variant set X^c for the variables group C is defined as the *union* of all variant sets X_i:

$$X^c = \bigcup_{i=1}^{x} X_i$$

   Alternative: Employ the *intersection* of the variant sets X_i, and keep only those groups where the resulting set of variants $X^c \neq \emptyset$:

$$X^c = \bigcap_{i=1}^{x} X_i$$

(b) Encoding groups as features for classification: Calculate *micro-average* for each variant $v \in X^c$ associated with the group $C = \{V_1, ..., V_n\}$, using data t:

$$mic(v, C) = \frac{\sum_{i=1}^{n} f(v, V_i, t)}{\sum_{i=1}^{n} \sum_{j=1}^{m} f(v_j, V_i, t)}$$

where $f(v, V, D)$ is defined as above (1), and $D = t$ is a given text.

5. Terminate the evaluation after a particular cut-off r yielded one single group containing the whole variables set. Such r, yielding a single group, corresponds to the *max. cut-off*, meaning that the dendrogram was cut at the root node.

Connection to the types of linguistic variables At this point, it is important to note the following connections between the grouping technique, its input and output, and the types of linguistic variables presented in Section 8.2:

1. *Absolute vs. Relative*

   The technique uses the concept of a *relative variable*, i.e., we represent each variable by a vector encoding the variants of this variable, employing the corresponding relative frequencies as values.

2. *Lexical vs. Grammatical*

   The technique is designed primarily for *lexical variables*. However, if in certain contexts grouping grammatical variables seems meaningful, the technique can be applied to them as well.[6]

[6]From the technical point of view, there is no difference in the procedure.

3. *Level of Granularity*

   (a) *FGV*: The actual items, i.e., the vectors representing original individual variables, serving as input to the technique

   (b) *CGV*: The single group obtained from the dendrogram at the *max. cut-off*, as described above in step 5 (see p. 115)

   (c) *IGV*: The various groups obtained from the dendrogram at the different cut-offs $<$ *max. cut-off*

8.4 Conclusions

In this chapter, we revised the notion of a linguistic variable, and based on the related research, we proposed a broader definition, which we consider more suitable in the context of this thesis. Furthermore, we suggested a taxonomy of linguistic variables, useful in this work. Finally, we presented a technique that can be used to abstract over individual variables by grouping those variables together which behave alike with respect to the classification label, i.e., the L1 of the learners, thus providing a more general perspective on the data.

In the following chapters of this part, we explore applying a variationist perspective to the task of NLI in more detail. We investigate our variationist approach from both, the quantitative and the qualitative perspectives using some selected variationist features, and discuss its advantages and limitations.

Connecting the findings and the research questions Regarding the particular five research questions in the focus of this thesis (see Section 1.3), the discussion in this chapter contributes to two of them as follows:

1. [VARIATIONIST-PERSPECTIVE]: In this chapter, we discussed the core conceptual issues in connection with a variationist perspective on a task such as NLI, and laid the ground for the implementation of a variationist approach in the context of this thesis. In particular, we revisited the notion of a linguistic variable, proposed a taxonomy of linguistic variables useful in the

given context, and presented a flexible technique for generating linguistic variables at different levels of abstraction.

2. [GENERAL-STRUCTURES]: We proposed a technique that is capable of abstracting over individual features to classes based on the variationist gist. It offers a huge set options, which can be explored in terms of the quantitative as well as qualitative analyses.

Chapter 9

Quantitative Explorations of the Variationist Approach

9.1 Introduction

In this chapter, we explore several feature types under a variationist perspective, following the definition of the linguistic variable proposed in Section 8.1. The focus in this chapter is on the quantitative aspect. We investigate two particular sorts of variables which on the one hand, are related in that both are situated at the level of syntax, but on the other hand, largely differ with respect to the taxonomy presented in Section 8.2. We start with the exploration of the linguistic variables reflecting the *syntactic category realization* in Section 9.2, and then, expanding the work by Krivanek (2012)[1], we turn to the variables encoding the variation in the *verb subcategorization* context in Section 9.3

9.2 Syntactic Category Realization (CFGR)

In this section, we systematically explore lexicalized and non-lexicalized local syntactic features (CFG production rules) for the task of NLI. We investigate different types of feature representations, with the main interest on the relative performance of two representations inspired by a variationist perspective. The varia-

[1] See Section 7.3 for more details.

tionist features explored in this section are *grammatical variables*, in that they are related to syntactic categories (see Section 8.2). We investigate *relative* as well as *absolute* versions by employing corresponding variationist feature encodings. Further, we evaluate the performance of our system in single- and cross-corpus settings.

This section presents parts of our results reported in Bykh & Meurers (2014), extending this contribution by some new conceptual considerations and connections, elaborated in the context of this thesis.

9.2.1 Data

The research in this section employs two data sets used in Chapter 6 (see Section 6.3) in the context of the First NLI Shared Task (see Section 3.2), namely T11 only for single-corpus, and NT11 $\cup$ T11 for cross-corpus evaluations. The data splits follow the settings used for the *closed* and *open-1* tasks in the context of the First NLI Shared Task, namely:

- *Single-corpus*: T11 *train* $\cup$ *dev* sets for training, and T11 *test* set for testing
- *Cross-corpus*: NT11 set for training, and T11 *test* set for testing

9.2.2 Tools

For parsing the T11 and NT11 corpora we employed the *Stanford Parser* (Klein & Manning, 2002). For classification, we used the *L2-regularized Logistic Regression* from the LIBLINEAR package (Fan et al., 2008), which we accessed through WEKA (Hall et al., 2009). To obtain results for all feature representations which are comparable across the different settings we uniformly scale all values employing the *-Z* option of WEKA.

9.2.3 Features

In this section, we focus on the CFG production rules (CFGR) as syntactic features. CFG rules are the most basic and widely used local syntactic units modularizing the overall syntactic analysis of a sentence. We parsed the T11 and NT11

corpora and extracted all CFG rules from the T11 and NT11 training sets. On this basis we defined the following tree feature types:

1. $CFGR_{ph}$: Only *phrasal* CFG production rules excluding all terminals
   - S $\rightarrow$ NP VP, NP $\rightarrow$ D NN, ...
2. $CFGR_{lex}$: Only *lexicalized* CFG production rules of the type *preterminal* $\rightarrow$ *terminal*
   - JJ $\rightarrow$ *nice*, JJ $\rightarrow$ *quick*, NN $\rightarrow$ *vacation*, ...
3. $CFGR_{ph \cup lex}$ = $CFGR_{ph} \cup CFGR_{lex}$ (i.e., the union of the above two)

In order to obtain a more comprehensive overview, we explore four different feature representations. First, we utilized two standard representations, namely:

1. *frequency-based (freq)*: A representation where the values are the raw counts of the occurrences of the rule in the given document.[2]
2. *binary (bin)*: A representation which only indicates whether a rule is present or absent in a given document.

Second, we focus on the exploration of two options that take as starting point the observation that CFG rules with the same left-hand side category represent different ways to rewrite that category. So in a sense, under a top-down perspective, there is a choice between different ways of realizing a given category. This suits the general logic of a variationist perspective and, in particular the definition of the linguistic variable proposed in Section 8.1. Under a variationist perspective, producing one of the variants of a given variable also means not choosing the other variants of that variable. So it is this grouping of observations that we want to take into account in terms of encoding local trees as features when we interpret the mother category as the variable to be realized and the different CFG rules with that left-hand side as variants of that variable. This results in two feature representations:

[2]Note that since we uniformly scale all values employing the -Z option of WEKA for better comparability, the *freq* feature representation based on the raw frequencies in essence also becomes normalized. This is particularly relevant in the context of the cross-corpus evaluation, where raw frequencies are particularly questionable given highly variable text sizes.

1. *simple variationist (var$_s$)* feature representation

2. *weighted variationist (var$_w$)* feature representation

The *var$_s$* encoding directly implements the *relative variable* idea, whereas *var$_w$* is a possible realization of the *absolute variable* logic, discussed in Section 8.2.1. More formally, the *var$_s$* and *var$_w$* frequency normalizations for each variant v from the set of variants V realizing a particular variable out of the set of variables $\overline{V}$ is defined as follows:

$$var_s(v \in V) = \frac{f(v)}{F(V)}$$

$$var_w(v \in V) = var_s(v) \cdot w(V)$$

Here, $f(v)$ yields the frequency x of a particular variant v, $F(V)$ is the sum over the frequencies of all variants v realizing the variable V, and $w(V)$ is the weight for the variable V:

$$f(v) = x$$

$$F(V) = \sum_{v \in V} f(v)$$

$$w(V \in \overline{V}) = \frac{F(V)}{\sum_{i=1}^{n} F(\overline{V}_i)}$$

The weighting applied in *var$_w$* takes into account the frequency proportion of each variable V in the overall variables set $\overline{V}$, assigning higher weights for more frequent variables. Since $F(V)$ cancels out in the definition of the *var$_w$* encoding, mathematically it reduces to normalizing each variant by the sum of the frequencies over all variants across all variables, i.e., to the relative frequency

of each variant v with respect to the set of all variants across all variables $\overline{V}$. As pointed out above, this way to encode variants can be attributed to the logic of a *absolute variable*[3] and thus, it is still a reasonable option in the context of a variationist approach. Eventually, there are different ways to express particular intentions or facts in terms of the linguistic form. In particular, it also concerns certain realizations of different syntactic categories explored in this section – E.g., it is well possible to express the same facts using a NP, or a VP as linguistic form, etc. (cf. Weber, 2012, 2014). The var_w encoding seems capable of capturing such phenomena.

9.2.4 Results

Single- vs. cross-corpus results The results for the *three feature types* using the *four different feature representations* are presented in Table 9.1. The chance baseline for the given data setup is 9.1%. There are big accuracy differences between the single- and cross-corpus settings despite very similar feature counts. The drop for the cross-corpus settings is roughly around $\frac{1}{2}$ compared to the single-corpus settings. This outcome is in line with previous results based on the same data setup using a wide range of features (see Section 6). It again confirms the fact that in general obtaining high cross-corpus results remains challenging in NLI.

Best feature type The $CFGR_{lex}$ feature type clearly outperforms the more abstract $CFGR_{ph}$ feature type, yielding up to 28% difference in accuracy for the single-corpus and up to 9% for the cross-corpus settings. The results show that the lexicalized feature type $CFGR_{lex}$ is useful in both, the single-corpus as well as the cross-corpus settings. It combines syntactic and lexical information, such as the fact that a given token with a particular POS is used, e.g., the token *can* being used as a *noun* in *There is a can of beer in the fridge* instead of as the more frequent *modal verb* use in *He can dance*. Note that this is different from using word and POS uni-grams as features, where the relevant connection is lost. In both the T11 data, which is topic balanced, for single-corpus evaluation and the very

[3]Eventually, we normalize the frequency of each CFGR as variant by the frequency of all rules in the text, which is comparable to simply normalizing by the overall word frequency in the text, suggested for encoding absolute variables (see Section 8.2.1).

features	single-corpus (sc): T11 training				
	freq	*bin*	var_s	var_w	feat. #
$CFGR_{ph}$	**50.0%**	44.3%	48.5%	49.8%	14,713
$CFGR_{lex}$	75.7%	72.5%	71.0%	**76.9%**	83,402
$CFGR_{ph \cup lex}$	78.2%	73.6%	75.4%	**78.8%**	98,115

features	cross-corpus (cc): NT11 training				
	freq	*bin*	var_s	var_w	feat. #
$CFGR_{ph}$	21.3%	22.9%	26.3%	**27.7%**	15,253
$CFGR_{lex}$	26.7%	32.0%	28.8%	**36.8%**	78,923
$CFGR_{ph \cup lex}$	28.3%	34.3%	32.6%	**38.8%**	94,176

Table 9.1: Results for the $CFGR$ feature types based on the standard T11 *test* set

heterogeneous NT11 data containing a wide range of topics for cross-corpus evaluation, we obtained consistently better results for $CFGR_{lex}$ than for $CFGR_{ph}$. Some syntactic rules including lexical information thus seem to generalize well across topics. This further supports previous findings, stressing the high value of lexical features for NLI (see Part II). Combining $CFGR_{ph}$ and $CFGR_{lex}$ into $CFGR_{ph \cup lex}$ gives an additional boost in performance.

Best feature representation There are clear differences in Table 9.1 between the results for the four feature representations. var_w yields the best accuracies in five out of six settings, across different feature types and corpora.

The results show that WEKA-normalized raw frequencies such as *freq* yield the worst results in a cross-corpus setting but perform very well single-corpus, which is in line with the assumption that raw frequency features do not generalize well. Applying some post-hoc normalization did not change the situation. In our experiments, the performance of *freq* in a cross-corpus setting is up to 10% worse than what is yielded by var_w, despite comparable single-corpus performance. *freq* also consistently performs worse than var_s in the cross-corpus setting, despite outperforming var_s single-corpus.

Using binary features (*bin*) yields better results cross-corpus than *freq*, whereas in the single-corpus setting it is the other way round. The abstraction introduced by the binary feature representation thus shows a positive effect in terms of the

capability of the features to generalize to other data sets.

For the abstract $CFGR_{ph}$ features, *var*$_s$ performs better than *freq* or *bin* in the cross-corpus setting. The fact that the *var*$_w$ is performing consistently better than *var*$_s$ shows that weighting is important, and suggests that absolute variables can be of high relevance in terms of performance. Thus, incorporating the insight from VS is not only conceptually interesting as a theoretical perspective, but also provides quantitative advantages.

CFGR categories as variables As mentioned above, the best performance is achieved by combining $CFGR_{ph}$ and $CFGR_{lex}$ into the $CFGR_{ph \cup lex}$ feature type using the weighted variationist feature representation *var*$_w$, which follows the logic of an absolute linguistic variable. Thus, we focus on that feature type and explore it more in depth.

For this, we split the overall *var*$_w$ normalized $CFGR_{ph \cup lex}$ feature set by the individual variables, i.e., the different *mother nodes*. Then, we trained separate models, each consisting of features encoding the different variants, i.e., the different realizations in which a given mother node can be rewritten. Our aim is to investigate the accuracy of the individual variable-based models and their contribution to the overall performance.

Figures 9.1 and 9.2 depict the single-corpus (sc) and cross-corpus (cc) accuracies yielded by each individual variable-based model. For presentation reasons, the results are shown separately for the $CFGR_{ph}$ and the $CFGR_{lex}$ subsets.

The $CFGR_{ph}$ results in Figure 9.1 show that a small subset of variables performs relatively well. Most of the models perform poorly, yielding accuracies close to the chance baseline. The best performing variables are essentially the main phrasal categories, such as S, NP, VP, PP, ADJP, ADVP and SBAR.

The results for the $CFGR_{lex}$ in Figure 9.2 show a similar pattern. There is a subset of variables which perform relatively well, usually models based on the main POS categories, such as the nominal (NN) and verbal (VB) categories as well as adjectives (JJ), prepositions (IN) and adverbs (RB). Some punctuation marks also seem to play a role. The rest of the models yields accuracies around the chance baseline. This might be due to data sparsity given that the main POS categories also are the most frequent. But those main categories also have the

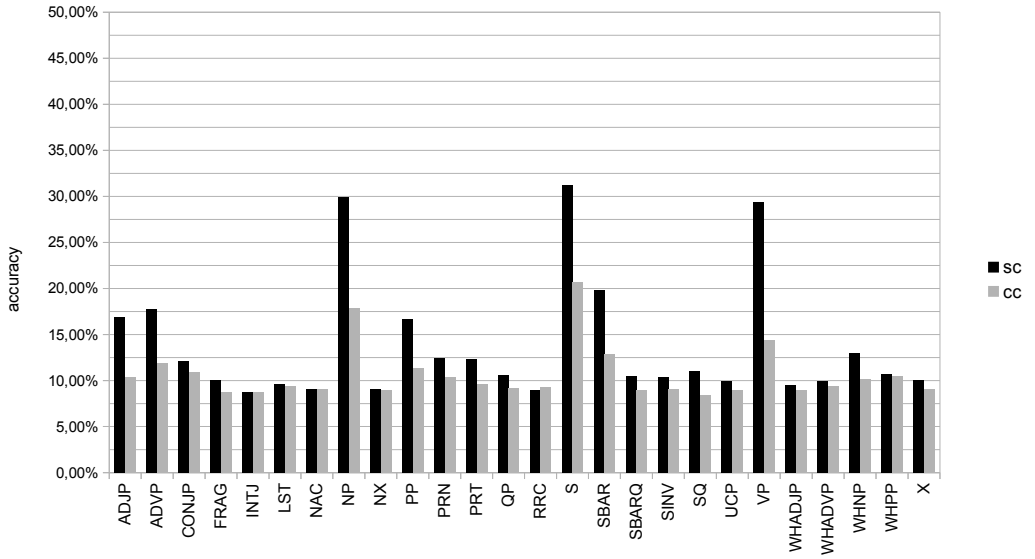

Figure 9.1: Accuracy for the individual $CFGR_{ph}$ variable based models, var_w

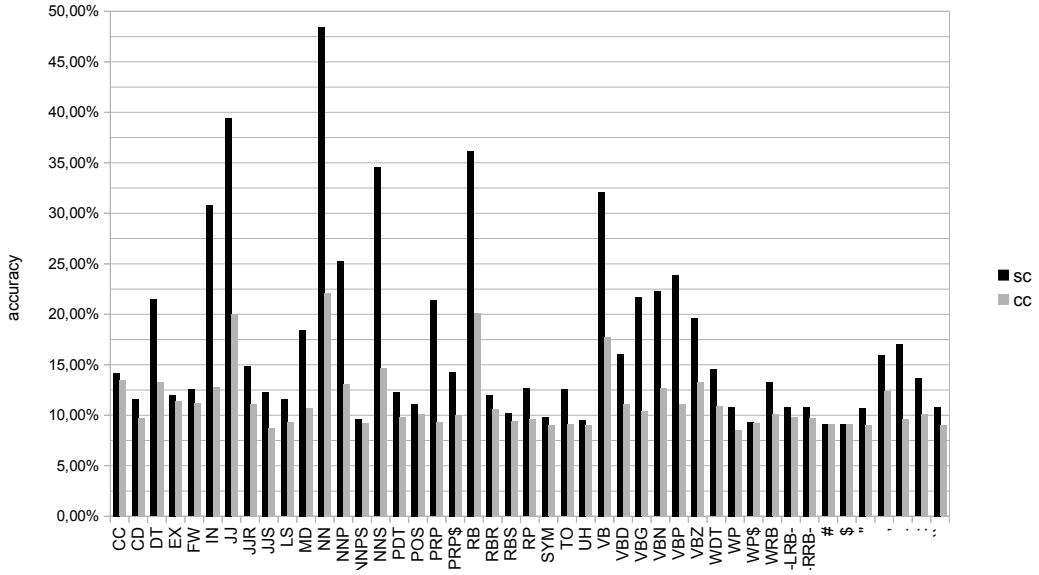

Figure 9.2: Accuracy for the individual $CFGR_{lex}$ variable based models, var_w

highest number of variants through which they can be realized. The good performance of the models for the variables with the highest number of variants thus supports the assumption that the choice of one of the realization options of a given category is influenced by the L1.

9.2.5 Conclusions

In this section, we systematically explored *non-lexicalized* and *lexicalized CFG production rules (CFGR)* as features for the task of NLI using both, single-corpus as well as cross-corpus settings. Including lexicalized CFG rule features clearly improved the results in both setting so that it seems worthwhile not to discard them a priori, which was usual in previous research.

Pursuing a variationist perspective to CFGR feature representation by modelling the features as *relative* and *absolute grammatical variables*, resulted in improved performance. Moreover, this supported an in-depth exploration of the contribution of the different variables and variants, as well as of the value of local syntactic features for NLI in general. Training a separate classifier for each variable supports a qualitative discussion of the categories reflecting the choices made by the learners with different L1s. Further research suggests that using such separate models can also provide quantitative advantages by facilitating high-performing ensemble setups – A point which we explore in detail in Part IV. The findings support our hypothesis that the choice of one of the realization options of a given syntactic category is influenced by the L1.

9.3 Verb Subcategorization (Verb Subcat)

In this section, we follow Krivanek (2012) and Meurers, Krivanek & Bykh (2014) and further extend it by systematically exploring different versions *verb subcategorization (subcat) features* under a variationist perspective for the task of NLI. We utilize dependencies as basis for feature generation.

On the one hand, the proposed feature type is well-suitable for our variationist approach, on the other hand it is also well-motivated in related SLA research (Tono, 2004; Callies & Szczesniak, 2008; Stringer, 2008), which makes it one of the best candidates for further explorations in the given context. Following the taxonomy presented in Section 8.2, we model the features as *relative lexical variables* by considering the distinct verb lemmas as variables and the different subcat patterns as variants realizing these variables. Our focus is on investigating a wide range of options regarding the *level of granularity* as introduced in Sec-

tion 8.2.3. Thus, we explore using FGVs, CGVs, and first of all different sets of IGVs, utilizing the *label-informed feature grouping* technique as proposed in Section 8.3.[4] To keep the results comprehensible and the number of experiments at a feasible level, in this section we perform only single-corpus experiments using the standard T11 data. Parts of this study are published in our contribution Bykh & Meurers (2016).

9.3.1 Data

The research in this section employs the T11 data introduced for the First NLI Shared Task (see Section 3.2) and used in in Chapter 6 (see Section 6.3). The data split follows the setting used for the *closed* task in the context of the First NLI Shared Task, namely: T11 *train* ∪ *dev* sets for training, and T11 *test* for testing.

9.3.2 Tools

We utilized the *MATE tools*[5] (Björkelund et al., 2010) for data preprocessing (tokenization, lemmatizing and POS-tagging) and the *MATE dependency parser* (Bohnet, 2010) to identify the arguments of a verb realized in a sentence, i.e., the subcategorization frame that was realized. For hierarchical clustering we employed *WEKA* (Hall et al., 2009). To process the resulting dendrograms we used the *Libnewicktree*[6] tree parser. Finally, classification was carried out using *L2-regularized Logistic Regression* from the *LIBLINEAR* package (Fan et al., 2008) accessed through WEKA, employing the *-Z* (normalization) parameter as in Section 9.2 for uniformity and better comparability of the results in this section.

9.3.3 Features

Simple vs. complex features The hypothesis we are testing is whether writers with different L1s prefer different subcat variants. To systematically explore the potential benefits of feature grouping, we start with *simple features*, where every

[4] See the note on p. 116 regarding the connection between the level of granularity for linguistic variables and the proposed grouping technique.

[5] `https://code.google.com/p/mate-tools`

[6] `https://github.com/cjb/libnewicktree`

variable, i.e., verb lemma, is considered separately. These features correspond to *relative lexical variables* at the *fine-grained* level (FGVs), following the taxonomy in Section 8.2.

We then infer sets of *complex features*, i.e., sets of various groups of variables using the proposed label-informed feature grouping-technique, abstracting from individual verb lemmas to classes of verbs (see Section 8.3). These features correspond to *relative lexical variables* at various *intermediate-grained* (IGVs) levels plus the *coarse-grained* level (CGVs), following the taxonomy in Section 8.2.[7] In contrast to Krivanek (2012) and Meurers, Krivanek & Bykh (2014), who group verb lemmas realizing *exactly the same* subcat variants, the technique we propose makes it possible to systematically explore a range of different groupings of lemmas based on the *similarity* of the realized subcat variants, and to take into account the L1 of the learners, as described in Section 8.3.[8]

In the following paragraphs, we describe the feature engineering procedure for the *simple features*. As discussed above, the *complex features* are obtained by applying the proposed label-informed feature grouping technique to the set of simple features.

Feature generation We dependency parsed the data and extracted the corresponding argument realization patterns, i.e., the realized subcat variants, for all verbs occurring in the data. We consider the following labels as arguments:

- *sbj*: subject
- *lgs*: logical subject
- *obj*: (in)direct object or clause complement
- *bnf*: benefactor in dative shift
- *dtv*: dative in dative shift

[7] See p. 116 for more details on the connection between the grouping technique and the level of granularity for linguistic variables.

[8] More formally, a *complex feature* is constituted by a group of verb lemmas C and the corresponding set of subcat variants X^c as described by step 4a in the overall grouping algorithm presented in Section 8.3.3, p. 114. All feature values are calculated using *micro average* as exemplified in step 4b of the overall grouping algorithm (Note that the simple features constitute a special case of the complex features, where the group C consists of a single variable, i.e., $C = \{V_1\}$).

- *prd*: predicative complement
- *oprd*: object complement
- *put*: locative complements of the verb put
- *vc*: verb chain

Utilizing the verb lemmas with their extracted subcat variants, we generated features, such as:

- *believe*_sbj
- *believe*_sbj_obj
- *may*_sbj_vc
- *put*_sbj_put
- *help*_sbj_obj
- *make*_sbj_obj_oprd, etc.

Feature reduction We performed the following three feature reduction steps due to some conceptual and practical considerations:

- **Reduction step 1 (RS1)**: Verbs as features are rather rare. In order to reduce data sparsity issues, at this point we opted for ignoring the *different permutations* of arguments within a subcat variant. This step reduced the number of distinct variants from 355 to 218.
- **Reduction step 2 (RS2)**: Some of the subcat variants are still rather specific and unlikely to occur frequently enough in the data. Some of them also suffer from tagging or parsing errors. So, in a second reduction step we grouped all *argument labels into three coarse-grained classes* in order to get more general patterns and to cope with data sparsity:
  - $\{sbj, lgs\} \rightarrow s$ (subject)
  - $obj \rightarrow o$ (object)
  - $\{bnf, dtv, prd, oprd, put, vc\} \rightarrow x$ (rest group)

The number of distinct subcat variants reduced from 218 to 48. Based on the three labels used here, we refer to these features as *SOX*. Applied to the examples listed above, they are of the following form:

  - *believe*_sbj → *believe*_s
  - *believe*_sbj_obj → *believe*_s_o
  - *may*_sbj_vc → *may*_s_x
  - *put*_sbj_put → *put*_s_x
  - *help*_sbj_obj → *help*_s_o
  - *make*_sbj_obj_oprd → *make*_s_o_x, etc.

- **Reduction step 3 (RS3)**: The last reduction step is conceptually different from the first two. It is based on theoretical considerations in connection with the variationist perspective, which we want to push further in this context: We are interested in the linguistic choices made by a speaker. If there is only a single variant for using a verb, we cannot observe a *choice* being made. We therefore dropped all features for verb lemmas that only occur with a single subcat variant in the training data.[9] That reduced the number of distinct verb lemmas from originally 11,401 to 3,785.

Feature reduction clearly also means a loss of potentially indicative subcat information. Eventually, there is a trade-off between reducing data sparsity or enforcing some theoretical concepts and keeping the feature set as large and as specific as possible to retain all potentially useful information. In the following we explore a range of options in connection with the first two feature reduction steps. The third step which, as pointed out, is based on important conceptual considerations, is retained for all experiments in Section 9.3.

9.3.4 Results

First, we explore in detail the most abstract feature type version, i.e., *SOX* (based on a reduced argument label set as explained above), ignoring the different argument label permutations. It means that we use features based on all three feature

[9]This reduction step is also the last in the logic of the algorithm. This means that here we consider the verb lemmas in connection with the 48 (and not the original 355) subcat variants, obtained after performing the previous two feature reduction steps.

reduction steps introduced in Section 9.3.3 (see p. 129). In the following, we refer to this particular feature version as *SOX-NoP*.

Second, we investigate the effect of making the features stepwise more specific, i.e., extending them by some additional information. On the one hand, we explore omitting the conceptually dispensable reduction steps, i.e., RS1 and RS2. On the other hand, we investigate incorporating additional linguistic information (via POS) into the feature structure. The five feature versions explored in this section are summarized in Table 9.2.

id	feature type	RS1	RS2	RS3	+POS
1	SOX-NoP	+	+	+	-
2	SOX-P	-	+	+	-
3	FULL-NoP	+	-	+	-
4	FULL-P	-	-	+	-
5	SOX-POS-NoP	+	+	+	+

Table 9.2: Different subcat feature versions explored in this section. *RS* refers to the corresponding feature reduction steps discussed in Section 9.3.3. The parameter *+POS* depicts if the argument labels were extended by the POS-tag of their heads.

Results for *SOX-NoP*

For the *SOX-NoP* feature type, we obtain 3,785 distinct verb lemmas (variables) with 14,389 subcat variants as individual features. That means that on average there are roughly four subcat variants per variable.

An overview of the results is presented in Figure 9.3. On the x-axis, the leftmost point, which is marked "s", corresponds to using only *simple features* (FGVs).[10] With increasing x-values, we go up in the dendrogram to obtain groups of verb lemmas, i.e., *complex features* (various IGVs, and the CGV at the root node). The x-values are the branch length cut-offs applied to the dendrogram using a step of 0.1 as discussed in Section 8.3. The y-axis represents the classification accuracy on the test set, using the training-test split described in Section 9.3.1 and the classifier spelled out in Section 9.3.2. The chance baseline is 9.1%.

[10]In other words, every verb lemma with the corresponding subcat variants is considered separately, i.e., no clustering.

Model with simple and complex features ([s/c]): This is the basic setting using simple and complex features. The figure shows that 44.5% at point "s" is the highest accuracy, so feature grouping does not provide a quantitative edge. For settings incorporating complex features, the best result is 44.2%, obtained for the cut-off 0.3. The clustering technique groups verb lemmas in terms of the proportion of their subcat variants. The particular verb lemmas, i.e., surface forms, which are part of a group, are not by themselves encoded in the feature space any more. For complex features the classifier therefore does not have an access to some potentially highly indicative surface properties. Even different misspellings of the verbs can be indicative of the L1 – distinctions that the proposed grouping method counting variants glosses over.[11] The disadvantage of loosing indicative surface information seems to outweigh the potential advantage of the generalization in

[11] E.g., we discovered that some of the clusters contained misspelled versions of the same verbs, such as *commmunicate/communicat*, *tounderstand/ubderstand*, *exaggrate/exxagarate*, etc.

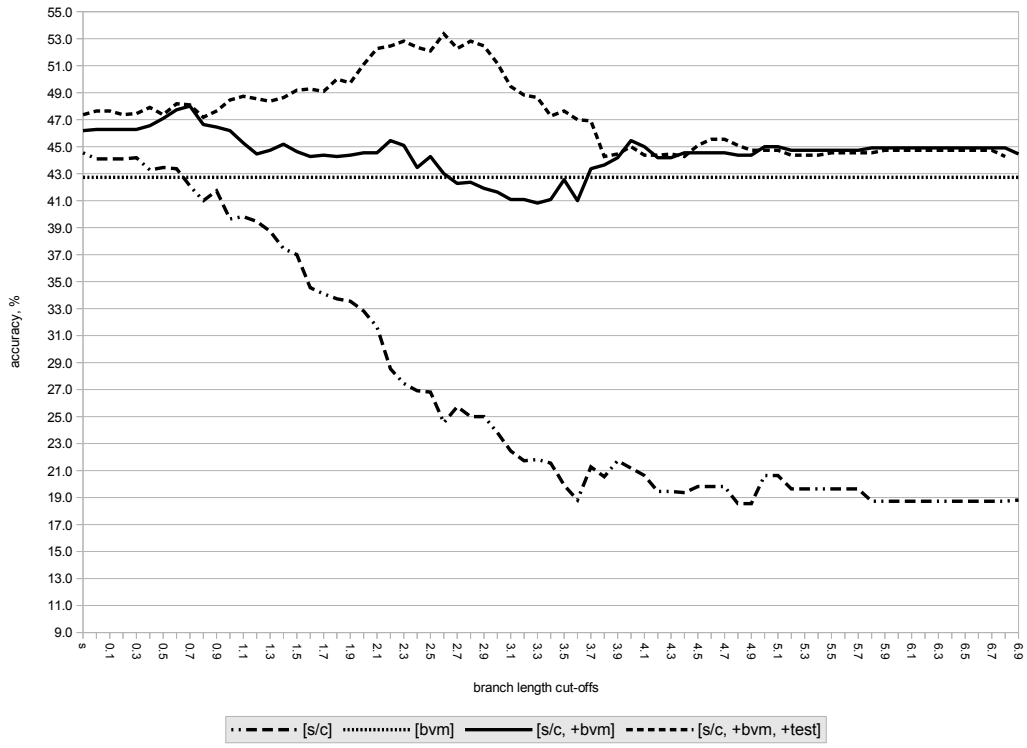

Figure 9.3: Classification accuracy employing the *SOX-NoP* feature type

terms of reducing data sparsity. We can validate that assumption by creating a *binary verb lemma model* and then combining it in a model with the simple and complex features.

Binary verb lemma model ([bvm]): To be able to identify the contribution of the subcat variants of individual verbs and groups of verbs, we need a way to separately quantify the information provided by the verb lemma itself (i.e., the presence of the variable, as separate from the choice of variants). In the [*bvm*] model, we thus only encode the presence/absence of the 3,785 verb lemmas for each text. The accuracy for that model is 42.7%. We included that result as a line in Figure 9.3 to visualize the performance relative to the other two models which make use of subcat pattern information. Comparing the other models to that one shows the benefits of incorporating the subcat variants as features. Indeed, the other curves (discussed below) show better results, confirming that subcat features are useful for NLI.

Binary verb lemma combined with simple and complex features ([s/c, +bvm]): To validate our assumption regarding the role of surface properties, we tried a combined setup, where the [*s/c*] setting was used in combination with the binary verb lemma model [*bvm*] described above. We assume the [*bvm*] model to restore the surface information lost due to the generalization, which should improve the classification performance. Indeed, the classification performance increased compared to the basic [*s/c*] setting. For simple features, the accuracy is 46.2%, and thus 1.7% higher. Including [*bvm*] therefore is beneficial even for settings not involving clustering. For settings including complex features, the difference depends on the actual cut-off. The best performance is 48.0% obtained using the cut-off 0.7, where the technique yielded 273 complex and 3016 simple features. The corresponding feature distribution is depicted in Figure 9.4. The difference between the best complex feature results is 3.8%. In most of the cases, the difference is even much higher, as seen when comparing the [*s/c*] and the [*s/c, +bvm*] curves at corresponding cut-offs in Figure 9.3. That result supports our assumption regarding the role of the surface properties. It is also supported by the shape of the [*s/c, +bvm*] curve only. The best model using complex features (cut-off 0.7)

outperforms the model solely based on simple features ("s") by 1.8%. Thus, when built on top of a surface-based model such as [*bvm*], the proposed grouping and generalization technique shows practical advantages in terms of accuracy. The findings confirm the hypothesis that in general learners with different L1s seem to prefer different subcat patterns.

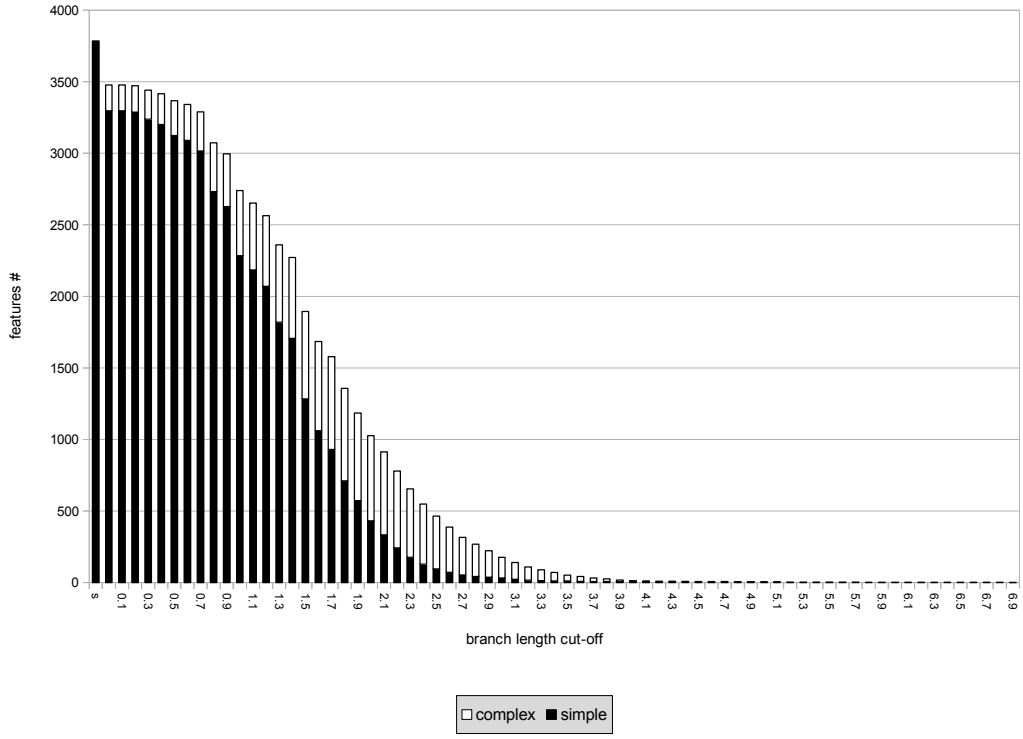

Figure 9.4: Distribution of simple and complex features for different cut-offs

Discussion Features based on verbs and the subcat patterns they realize are relatively sparse.[12] So we wanted to explore what would happen if we had a better coverage of the verbs that actually occur in the test set. We therefore performed another experiment, in which we used the whole corpus for feature grouping. The results for the setting are shown as **[s/c, +bvm, +test]** in Figure 9.3. Since this

[12]Cf. the discussion on data sparsity in the context of syntactic transfer in Odlin (2003, p. 440).

includes the test data set for feature generation, this experiment only serves to identify a performance ceiling for the clustering approach given this corpus. It helps to identify and study verb lemma clusters of different sizes under ideal circumstances in terms of the coverage of the verbal subcat variants.

The number of verb lemmas in this setup slightly increased from 3,785 to 4,019 and the number of the distinct subcat variants from 48 to 49. The best accuracy is 53.4% using the cut-off 2.6, thus 5.4% higher than without including the test data in the feature clustering procedure.

Comparing the curves for [*s/c, +bvm*] and [*s/c, +bvm, +test*] shows that at the cut-off 0.7, we obtain essentially the same best accuracy up to that cut-off point using solely the training set information for clustering. Here, using the complete verb subcat information for clustering does not seem to significantly enhance the cluster informativeness. Hence, for the given cut-off range the system seems to generalize in a reasonable way.

Second, the results suggest that having bigger clusters (which are obtained at higher cut-offs, cf. Figure 9.4) can indeed be more informative than utilizing only relatively small clusters usually obtained by using lower cut-offs: For the [*s/c, +bvm*] setting, the best accuracy of 48.0% was obtained at a relatively low dendrogram cut-off (0.7), resulting in using rather few small groups and a lot of clusters containing a single verb (273 complex and 3016 simple features); whereas for the better performing [*s/c, +bvm, +test*] setting, the cut-off for the best accuracy of 53.4% was much higher (2.6), resulting in using bigger groups and only few clusters containing a single verb (344 complex and 76 simple features).

Finally, the results show that the training and test sets are not sufficiently similar, or the training data is not sufficiently large for this rather sparse feature type, limiting the performance.

Relative performance In comparison with previous research, the presented automatic feature grouping technique outperforms the approach by Meurers, Krivanek & Bykh (2014), where only verb lemmas with equal subcat variant sets constituted a group (alternation). A replication of this approach using the SOX-NoP feature type yielded 38.7%, and after adding the *[bvm]* model, 44.4% – This is 3.6% lower than the best result in [*s/c, +bvm*] (cut-off 0.7) presented above.

Exploring Different Versions of Verb Subcat Features: Results for *SOX-P*, *FULL-NoP*, *FULL-P* and *SOX-POS-NoP*

In addition to the *SOX-NoP* feature type explored above in very detail, here we investigate how the performance is affected by making the subcat features more specific. In particular, we investigate settings omitting the proposed feature reduction steps as introduced in Section 9.3.3[13], and explore the effect of enriching the features by more linguistic information. On the one hand, this increases data sparsity, on the other hand more potentially indicative information is provided to the classifier. The results are presented and discussed in the paragraphs below, each representing separate experimental setups employing different feature versions, corresponding to the ids 2-5 as presented in Table 9.2. We use the *SOX-NoP* feature type as reference in our discussions.

***SOX-P*: Reduced argument labels set including permutations** In this first additional experiments set on verb subcat features, we omitted RS1, which means that we used *SOX* features, considering different permutations of the argument labels as separate variants. This resulted in using 108 distinct subcat variants identified in the training data. The number of distinct verb lemmas showing at least two different subcat realizations increased from 3,785 to 3,789, and the number of subcat variants as features increased from 14,389 to 15,249 compared to *SOX-NoP*, which differ only by ignoring the argument label permutations in the features generation process. Thus, the general form is the same as for the *SOX-NoP* feature type, with the difference that, e.g., *believe*_s_o and *believe*_o_s would constitute two distinct features.

The results are presented in Figure 9.5. The different curves correspond to the settings described above in the exploration on the *SOX-NoP* feature type, and they show a similar pattern. The best result of 47.3% (cut-offs 0.0 to 0.2) is yielded by the setting [*s/c, +bvm*] combining subcat features and the binary verb lemma model, the [*s/c*] setting using only subcat features is the second best, and the [*bvm*] setting, which does not use any subcat information but only verb lemma features,

[13]We experiment with omitting the first two feature reduction steps RS1 and RS2. Note that RS3 is retained for all settings due to conceptual considerations in connection with the variationist perspective as discussed in Section 9.3.3.

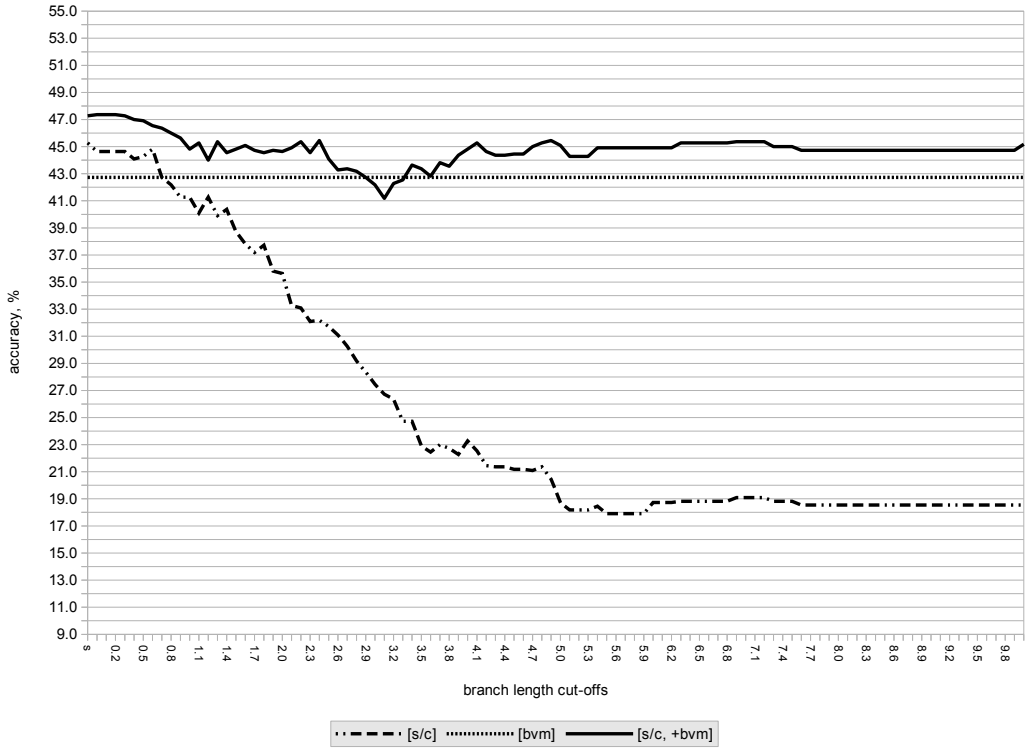

Figure 9.5: Classification accuracy for the SOX-P feature type

is the worst performing one. Different from *SOX-NoP*, there is hardly an improvement between the "s" and the settings incorporating grouping in [*s/c, +bvm*], i.e., it is only 0.1%. Moreover, the best performance of 47.3% is worse than for *SOX-NoP*, where we obtained a best accuracy of 48%. Thus, for the abstract features such as *SOX* reducing the variance by ignoring the different permutations of the argument labels proves reasonable from the quantitative point of view.

***FULL-NoP*: Full argument labels set ignoring permutations** In this second additional experiments set on verb subcat features, we omitted RS2, which means that we employed all original nine argument labels, resulting in using the 218 distinct subcat variants occurring in the training data (see Section 9.3.3). The number of distinct verb lemmas showing at least two different subcat realizations increased from 3,785 to 3,800, and the number of subcat variants as features increased from 14,389 to 15,952 compared to the *SOX-NoP* feature type, which, in

Figure 9.6: Classification accuracy for the FULL-NoP feature type

contrast, is based on a reduced argument label set. The generated features are of the following form:

- *believe*_sbj
- *believe*_sbj_obj
- *may*_sbj_vc, etc.

The results are presented in Figure 9.6. The different curves correspond to the settings described above in the report on the *SOX-NoP* feature type, and they show a similar pattern. The best result of 47.1% (cut-off 0.2) is yielded by the setting [*s/c, +bvm*] combining subcat features and the binary verb lemma model, the [*s/c*] setting using only subcat features is the second best, and the [*bvm*] setting, which does not use any subcat information but only verb lemma features, is the worst performing one. There are also some interesting differences compared to

SOX-NoP, e.g., the [*s/c*] setting shows a slightly improved performance (by 0.4%) at the cut-off 0.3 compared to the "s" setting, where no grouping is used – There was no improvement via grouping in this setting for *SOX-NoP*. Thus, the more specific subcat features seem to better compensate for the loss of some indicative surface traits encoded in the individual lemmas, which get glossed over by feature grouping. However, the increase in accuracy is rather marginal. The improvement between "s" and the settings incorporating grouping in [*s/c, +bvm*] is much smaller, namely, only 0.3%. In sum, the best performance of 47.1% is worse than with the *SOX-NoP* feature type, where we obtained a best accuracy of 48%. Thus, reducing the label set in order to obtain more general features as implemented by the *SOX-NoP* feature type proves reasonable in terms of performance.

***FULL-P*: Full argument labels set including permutations** In this third additional experiments set on verb subcat features, we wanted to observe what happens if we omit RS1 and RS2 at once.[14] This means that we employed all original nine argument labels and retained their different permutations in the feature generation process, which resulted in using all of the 355 distinct subcat variants occurring in the training data (see Section 9.3.3). The number of distinct verb lemmas showing at least two different subcat realizations increased from 3,785 to 3,803, and the number of subcat variants as features increased from 14,389 to 16,776 compared to *SOX-NoP*, incorporating all reduction steps. The general form is similar to *FULL-NoP* presented in the previous paragraph, with the difference that, e.g., *believe*_sbj_obj and *believe*_obj_sbj would constitute two distinct features.

The results are presented in Figure 9.7. Again, the different curves correspond to the settings described above in the report on the *SOX-NoP* feature type, and they again show a similar pattern. The best result of 48.2% (cut-off 0.3) is yielded by the setting [*s/c, +bvm*] combining subcat features and the binary verb lemma model, the [*s/c*] setting using only subcat features is the second best, and the [*bvm*] setting, which does not use any subcat information but only verb lemma features, is the worst performing one. Eventually, omitting all of the conceptually dispensable reduction steps resulted in an approx. 16% larger feature set with an improvement in accuracy by only 0.2% compared to the best setting incorporating

[14] In other words, we omit all *conceptually dispensable* reduction steps (see Section 9.3.3).

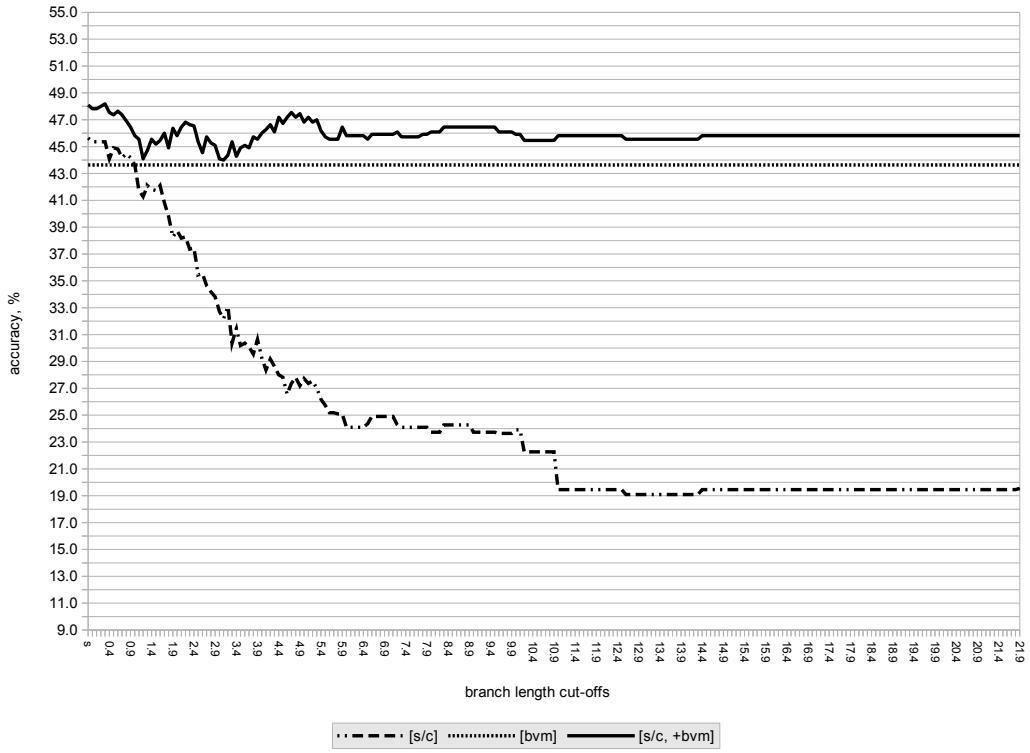

Figure 9.7: Classification accuracy for the FULL-P feature type

all of the reduction steps (cf. the [*s/c, +bvm*] setting for *SOX-NoP*). Thus, the proposed feature reduction notably decreased the feature space without significantly harming the accuracy. Interestingly, the difference between the "s" setting, where no grouping is employed, and the best performing grouping is only 0.1% – The same as for the *SOX-P* feature type discussed above. Note that for the more abstract *SOX-NoP* feature type this difference was 1.8%. Thus, again using more specific features limited the advantages of the grouping procedure.

***SOX-POS-NoP*: Extending the *SOX-NoP* feature type by POS** In this fourth and at the same time last additional experiments set on verb subcat features, we did not omit any reduction steps, instead we generated more specific features by incorporating additional linguistic information into the original *SOX-NoP* features, which constitute the most abstract and general version in our set. In particular, we extended the *SOX-NoP* features by the corresponding POS tags of the argu-

ment heads. For this we used the POS-tagger provided by the *MATE* tools (see Section 9.3.2). In order to decrease data sparsity, we employed a reduced POS tag set, where the different *PennTreebank* POS tags[15] for the main categories – namely, verbs, nouns, adjectives and adverbs – were represented by a single general tag – namely, *vb*, *nn*, *jj* and *rb* – correspondingly. The generated features are of the following form:

- *believe*_s+prp
- *believe*_s+nn_o+in
- *may*_s+prp_x+vb, etc.

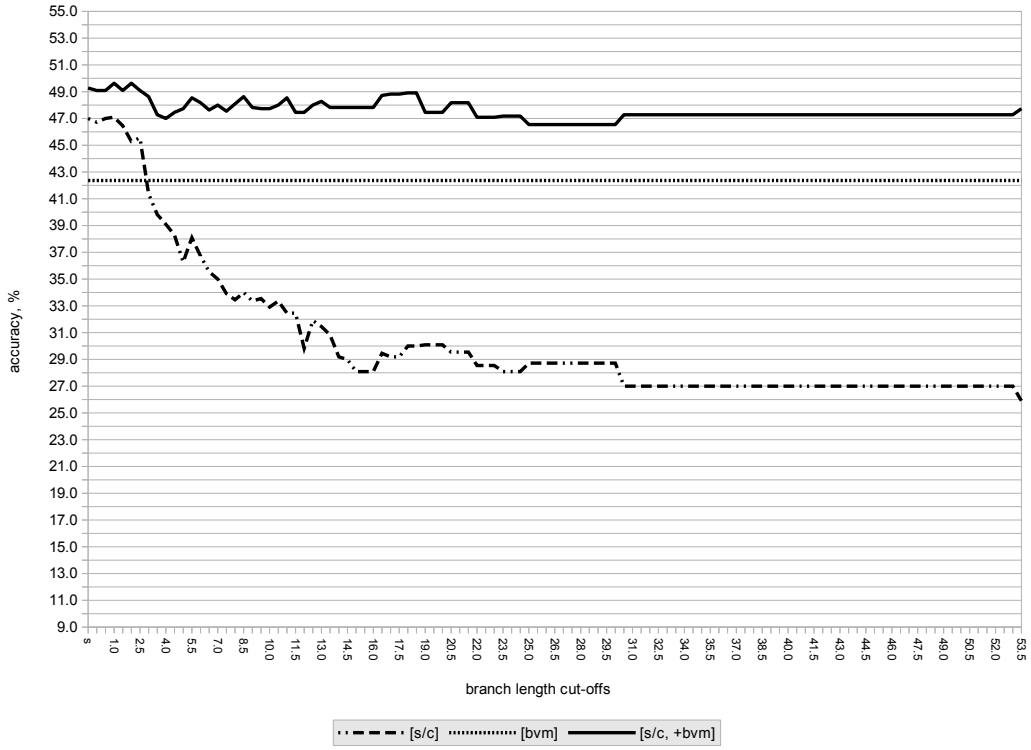

Figure 9.8: Classification accuracy for the SOX-POS-NoP feature type

[15] https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

The number of distinct subcat variants increased from 48 to 2,545, the number of distinct verb lemmas showing at least two different subcat realizations increased from 3,785 to 4,040, and the number of subcat variants as features increased from 14,389 to 36,618 compared to the original *SOX-NoP* feature type. The results are presented in Figure 9.8. The different curves again correspond to the settings described above in the report on *SOX-NoP*, and once again they show a similar pattern. The best result of 49.6% (cut-offs 1.0 and 2.0)[16] is yielded by the setting [*s/c, +bvm*] including subcat features and the binary verb lemma model, the [*s/c*] setting using only subcat features is the second best, and the [*bvm*] setting, which does not use any subcat information but only verb lemma features, is the worst performing one. In general, the results resemble the findings on *FULL-P* presented in the previous paragraph – Compared to *SOX-NoP*, the best accuracy is 1.6% higher, but it comes along with an increase in the feature counts of more than the factor two (35,298 vs. 16,841 features at the corresponding best cut-offs). At the same time feature grouping shows only marginal improvements (by 0.3%).

Discussion Applying the proposed feature grouping technique always increased the accuracy in the [*s/c, +bvm*] settings, which combine verb subcat features and the binary verb lemma model. However, we observe that whenever we employed features which were more specific than the *SOX-NoP* feature type, the quantitative advantages of the proposed feature grouping procedure decreased, yielding only marginal improvements. In the end, for a larger feature space with more specific features, the chance is higher that the classifier will identify some highly indicative variable-variant pairings. However, grouping glosses over the individual variables, so that such specific features are not directly accessible in the training procedure any more. Moreover, using a more specific feature set decreases the frequencies of the individual features, making them less reliable and thus potentially harming the clustering procedure, i.e., the quality of the resulting groups. In sum, the more specific the features, the less useful becomes the feature grouping. In other words, from the quantitative perspective, the proposed feature grouping

[16]The *max. cut-off* for this feature type is 53.4, which is relatively high. Note that for the *SOX-NoP* feature type it was only 6.9. In order to reduce the number of single experiments to a feasible extent, we used the cut-off step $s = 0.5$ (instead of the default 0.1) in this setup.

technique seems to be more advantageous if used in combination with more abstract and general linguistic features.

9.3.5 Conclusions

In this section, we modelled verb subcategorization features as *relative lexical variables* and explored the performance of different sets of variables at the *intermediate-grained* level, employing the *label-informed feature grouping technique* proposed in Section 8.3. It turned out that grouping features, i.e., utilizing different sets of intermediate-grained variables by considering different dendrogram cut-offs, indeed provides quantitative advantages, especially if used in connection with linguistic features at higher levels of abstraction. For more specific features, the potential advantages provided by grouping seem to decrease. We assume that this might be due to the following main reasons: On the one hand, grouping glosses over some potentially highly indicative cues; on the other hand, a more specific feature set leads to lower and thus less reliable frequencies of the individual features, potentially harming the clustering procedure, i.e., the quality of the groups. Using the *feature reduction* steps proposed in Section 9.3.3 seem to have mainly positive effects: Omitting RS1 and RS2 separately, showed a decrease in terms of accuracy; omitting both at once, and thus making the feature set much more specific, showed a marginal accuracy improvement of 0.2%, at the same time increasing the number of features by 16%. Thus, using the proposed feature reduction either increased the accuracy, or marginally decreased it, at the same time notably reducing the feature space. Further increasing the specificity of the features by incorporating the corresponding POS information of the arguments for the basic feature type (*SOX-NoP*), showed an improvement in terms of accuracy by 1.6%. However, at the same time the number of features increased by more than the factor two. Thus, it is well possible to improve the quantitative results by utilizing some linguistically more informed features, but it comes at the price of a significant increased in the model complexity. In sum, the high quantitative results support the hypothesis that learners with different L1s seem to prefer different subcategorization patterns in their non-native productions.

9.4 Conclusions

In this chapter, we explored a range of linguistic features under a variationist perspective, focusing on the quantitative aspect. We exemplified the taxonomy presented in Section 8.2, and showed for the particular area of syntax how linguistic variables of different types can be generated based on the principles of VS. Our findings suggest that using feature encodings based on the variationist perspective can indeed provide a quantitative edge, compared to simple frequency or binary representations. Furthermore, we applied the label-informed feature grouping proposed in Section 8.3 to suitable linguistic variables and explored the quantitative potential of the technique. It turned out that it can provide quantitative advantages, first of all, if applied to more abstract linguistic features and supported by surface models. It seems capable of generating reasonable groups of variables, and to optimize the models in terms of feature counts and data sparsity sensitivity.

The accuracies are all well above chance, ranging up to around 79% given a chance baseline of 9.1%. Thus, the quantitative results suggest that learners with different L1s, indeed seem to prefer different variants when realizing different linguistic variables explored in this chapter. In referring to the provocative question in the title of this thesis, if there is a real choice in the production of non-native output, or if the choice is predetermined by the L1, we can state the following: The accuracies are relatively high, so the choice seems to be definitely influenced by the L1. However, there are clearly many factors such as proficiency, social status, age or gender, etc., which certainly also influence the preferences to different extents, not clear at this point. Unfortunately, controlling for all of them seems hardly feasible with the data used for our explorations. Additional studies are required to further approach the answer to this highly thrilling question.

Finally, in this chapter, we conducted only some quantitative investigations. However, in order to advance the theoretical insight on L1-transfer, i.e., to contribute any valuable insight to the SLA research, it is clearly not sufficient to explore only the performance of the different systems. High performance is an important indicator for features, capturing potentially interesting differences in the language use by learners with different L1s. Nevertheless, identifying high

performing settings can only be the first step in the overall procedure of advancing the scientific insight. The next step is to employ appropriate methods and to spot and motivate particular interesting L1-transfer candidates. This is exactly what we target in the next chapter, where we turn to the qualitative exploration of different variationist features.

Connecting the findings and the research questions Regarding the particular five research questions in the focus of this thesis (see Section 1.3), the discussion in this chapter contributes to each of them as follows:

1. [VARIATIONIST-PERSPECTIVE]: In this chapter, we explored applying the variationist perspective to the task of NLI from the quantitative point of view. Our findings suggest that using a variationist view on features and feature encoding, can indeed enhance NLI systems and provide quantitative advantages in terms of classification accuracy.

2. [GENERAL-STRUCTURES]: We explored using the label-informed feature grouping technique proposed in Section 8.3. We showed that it is well-capable of abstracting from individual features to reasonable classes. The approach emphasizes how the underlying linguistic structure informs the classification label, reducing potential problems arising from idiosyncrasies and sparsity of individual features. It offers a huge set options to explore, and is well-capable of advancing the performance of NLI systems.

3. [LINGUISTIC-FEATURES]: In this chapter, we systematically investigated the performance of different linguistic features at the syntax level under a variationist perspective. The high accuracies support the previous research, suggesting that syntactic features seem to be of high relevance for the task of NLI, and in the context of L1-transfer in general. At the same time, our findings further confirm the high relevance of lexical traits for the task at hand (cf. Part II): The lexicalized version of the CFGR feature type outperformed the phrasal CFGR features, and the proposed label-informed feature grouping technique benefited from being supported by a lexical model.

4. [CROSS-CORPUS]: We compared the single- and cross-corpus performance for a subset of the features (see Section 9.2). As expected, the findings supported the previous research, showing lower accuracies if the systems are trained and tested on different corpora. Employing a lexicalized version of the CFGR feature type showed highest accuracies, further confirming our findings on the importance of lexical information for the cross-corpus performance in NLI (cf. Part II).

5. [MODEL-OPTIMIZATION]: The explored label-informed feature grouping technique seems to be an appropriate way for optimizing feature sets following the variationist logic. It is capable of automatically inducing more general features, which on the one hand, reduces the feature space, thus making the models more compact, and on the other hand, can improve the accuracy by overcoming the data sparsity of individual features.

Chapter 10

Qualitative Explorations of the Variationist Approach

10.1 Introduction

In Chapter 9, we focused on the exploration of a variationist approach to NLI from the quantitative perspective. Complementing our quantitative findings, in this chapter we turn to the qualitative explorations, showing how a variationist approach to NLI can contribute insight to the SLA research. We exemplify the following two core directions employing our method:

1. Discovering potentially interesting L1-transfer candidates in an automatic data-driven way

2. Testing specific (existing and new) hypotheses about L1-transfer

Our main goal is not to provide as many as possible individual findings on L1-transfer, but rather to show how the proposed method can serve as a practical tool for obtaining interesting and instructive qualitative findings, advancing the SLA insight – A new interesting and, in our opinion, important direction emerging in NLI, which seems worthwhile following and extending (Swanson & Charniak, 2013, 2014; Malmasi & Dras, 2014a). We exemplify our method mostly targeting *L1 German*. However, it is clearly possible to explore any other L1 of interest,

following the same general procedure.[1] This chapter includes some of our findings reported in Bykh & Meurers (2016), and further extends this work.

10.2 Evaluation Setup

Before we begin our analysis, some core issues have to be clarified. In the context of the qualitative explorations, the requirements differ from those in the quantitative part. In the following we discuss the differences and motivate our decisions made for the work presented in this chapter.

10.2.1 Data

Before we can start any investigations, the first question to clarify is what data to use for the qualitative analysis? In general, the more data can be employed, the more reliable the findings can be. However, especially in the context of the qualitative analysis, the actual quality of the data plays a central role. There are many different parameters, such as the proficiency or the prompts distribution, which can influence the outcomes. Unfortunately, it does not seem feasible to control for these parameters with our rather heterogeneous NT11 data set, consisting of a number of different corpora compiled using different guidelines (see Section 6.3). Thus, here we decided to focus on the most large, consistent corpus in our data, namely T11, introduced for the First NLI Shared Task (see Section 3.2). It was compiled using strict guidelines, shows a uniform structure, and is prepared specifically for the task of NLI, i.e., accounts for issues such as different character encodings, topic bias, etc. (see Blanchard et al., 2013; Tetreault et al., 2012). The data split follows the setting used for the *closed* task in the context of the First NLI Shared Task, namely: T11 *train* $\cup$ *dev* sets for training, and T11 *test* for testing. For now, we consider this data as the best option for our qualitative explorations. Verifying our findings using other corpora is an important research direction in terms of the future work.

[1] We include some raw data for all L1s in our data set in the different tables presented in Appendix A and Appendix B, with the aim to support further analyses beyond the discussions in this chapter.

10.2.2 Tools

We employ essentially the same tools as for our quantitative explorations in Section 9.3.2. In particular, we utilized the *MATE tools*[2] (Björkelund et al., 2010) for data preprocessing (tokenization, lemmatizing and POS-tagging) and the *MATE dependency parser* (Bohnet, 2010) to generate the different dependency-based features utilized in this study. For hierarchical clustering we employed *WEKA* (Hall et al., 2009). To process the resulting dendrograms we used the *Libnewicktree*[3] tree parser. Finally, classification was carried out using *L2-regularized Logistic Regression* from the *LIBLINEAR* package (Fan et al., 2008). Different from some of the performance oriented studies, we did not employ any post-hoc normalization of the feature vectors (see the discussion in Section 10.2.3).

10.2.3 Feature Choice and Encoding

The next question is what particular features to use for our qualitative explorations? On the one hand, in order to obtain reliable and general qualitative findings, it seems necessary to choose features that are as little as possible affected by data sparsity. On the other hand, the features should be linguistically interpretable. Moreover, our aim is, on the one hand, to cover functionally different cases, e.g., complements vs. adjuncts; on the other hand, to design features reflecting structures that are known to be instructive in the related VS and SLA research. Thus, in this chapter, we explore features implementing different linguistic variables which we consider to be among the most suitable under the given constraints.

While the original features, we utilized for the quantitative explorations in Chapter 9, i.e., syntactic category realization and verb subcat features, have been well-suitable for exemplifying the quantitative advantages of the approach, they do not seem to be best suitable for the qualitative explorations given the mentioned constraints, at least not in the original form. Regarding the verb subcat features, some of them are very specific and seem to be prone to data sparsity; others seem to be mostly a way too abstract to allow for a reasonable linguistic interpretation. The category realization features seem in general more suitable, but lack some

[2] `https://code.google.com/p/mate-tools`
[3] `https://github.com/cjb/libnewicktree`

interesting information, such as the functions. Finally, we opted for generating new features, which are related to the previously explored ones, or pose a mix of those, and in our opinion are generally better suitable for the qualitative analysis, namely features encoding *subject (pronoun) realization* and *nominal modification*. We introduce and discuss them in detail in Section 10.3.

Apart from that, we do not use any additional feature normalization techniques, employed in other parts of this thesis (see Section 9). It was shown that depending on how the features are normalized, it can have a positive effect on the NLI performance (see Jarvis et al., 2013). However, any feature transformation introduces additional complexity and can hamper the interpretation of the results. So, for simplicity and transparency reasons, we decided against applying any additional feature transformation steps in this chapter.

10.2.4 Feature Evaluation

Next, we have to clarify what feature evaluation method to use in order to identify patterns which might be instructive and valuable from the qualitative point of view?

There are different possible options. However, in order to keep a connection to other results in our work, to ensure a better comparability of the findings within this thesis, as well as a better interpretability and traceability of the outcomes, we opted for using the same algorithm for feature evaluation as was employed for performing the actual classification in most of the experiments in this thesis, namely, the *L2-regularized Logistic Regression (LR)* provided by the *LIBLINEAR* package (Fan et al., 2008). Moreover, logistic regression as method was already successfully employed for different analyses in the VS context before (Tagliamonte, 2012). In the following we suggest a simple evaluation approach, which we will use for our qualitative explorations in this chapter.

In particular, we train individual *one-vs.-rest* classifiers (Witten et al., 2011; Alpaydin, 2004) for the given 11 L1s as classes, and then extract and rank the weights (coefficients) w_i assigned by these individual classifiers to the particular features f_i. This allows to determine the contribution of the individual features to the classification decision, i.e., their predictive power. On the one hand, the sign

of the weight shows if the corresponding feature contributes to a classification decision toward the class of interest (positive class) or the *rest* class (negative class) as follows:

- $w_i > 0$ means that the corresponding feature f_i contributes a piece of evidence to a classification decision towards the class of interest (positive class).
- $w_i < 0$ means that the corresponding feature f_i contributes a piece of evidence to a classification decision towards the *rest* class (negative class).
- Finally, $w_i \approx 0$ means that the corresponding feature f_i has no distinctive power (i.e., the feature is not helpful for distinguishing between the given classes based on the given training data).

On the other hand, the magnitude of a weight shows the importance of the corresponding feature for the classification decision (cf. Alpaydin, 2004; Witten et al., 2011). Thus, we consider features which get assigned *indicative positive* w *values* in connection with a particular L1 as class, as potentially interesting *candidates* for further qualitative explorations in the context of L1-transfer. In the following, we define the criteria for selecting a particular feature f as an interesting candidate for our analysis.

Let L be the L1 of interest, W_f the weights distribution (consisting of the weighs for all L1s in the set) for a particular feature f, and $w_L \in W_f$ the weight for L in the given weights distribution for f. For selecting a particular feature f as one of the *most indicative* for the given L, the following criteria have to be matched:

1. $m(W_f) \leq 0.0$, where m is the *median* of the weights distribution W_f.
    - This ensures that at least half of $w \in W_f$ show values $w \leq 0$, for f to be selected. This prevents selecting features which are indicative for too many different L1s, and thus do not seem to be distinctive and interesting enough.
    - For our data incorporating 11 L1s, it means that only f indicative for *at most* five L1s can be selected, other f are dropped.

2. $w_L \in M = \{w \mid w \geq max(n, W_f)\}$, where $max(n, W_f)$ yields the top nth weight $w \in W_f$, and we use $n = 3$.

   - This ensures that w_L is among the top n weights $w \in W_f$, thus increasing the probability for selecting interesting f.
   - This parameter can be clearly varied depending on the strictness of the approach and the data. Setting $n = 1$ would mean that only features, where w_L is the highest weight in W_f are considered. However, this would significantly reduce the set of potentially interesting candidates. Thus, we opt for the more relaxed $n = 3$, which seems still reasonable with 11 L1s: w_L has to be among $\approx \frac{1}{3}$ of the highest weighted L1s in connection with f, for f to be selected.

3. $w_L > t$, using $t = 0.1$

   - Since matching the criteria 1 and 2 can still show weights $w_L \leq 0$, this criterion ensures some minimum indicative power for f to be selected.
   - This parameter can be clearly varied depending on the strictness of the approach and the data. For our evaluation we pick a rather low t, in order to avoid missing too many potentially interesting candidates.

Discussion Malmasi & Dras (2014d) suggest a similar method employing linear SVM weights. In particular, they compile and evaluate two ranked lists – One containing the positive weights, and another one containing the negative weights for a particular L1. Then the authors interpret the positive weights in terms of *overuse* and the negative weights in terms of *underuse* of certain patterns in connection with a particular L1. However, overuse and underuse is generally determined by notable or significant differences in the frequencies of particular language units (see Section 2.1). Yet, just the raw weights do no tell much about the actual frequencies or statistical significance. Thus, while the reported cases of potential L1-transfer are interesting and well-motivated, it is not ensured that the particular interpretation of the weights is conceptually sustainable in general. We believe that these two aspects, namely, valuable findings based on the classifier weights, and the particular interpretation in terms of overuse/underuse in the context of

SLA research should be considered as two separate issues. So, in applying our method, here we target the first aspect, i.e., we suggest to investigate high positive weights for particular L1s, which point to features indicative for these L1s, and thus can help identifying interesting L1-transfer candidates – no more, no less. Nevertheless, the method can well serve as a helpful first step in the identification of overuse/underuse patterns, which are clearly of interest for the research.

An alternative related approach would be using some standard feature selection techniques such as *Information Gain*, or dimensionality reduction methods such as *PCA* (Kubat, 2015; Witten et al., 2011; Alpaydin, 2004) for the feature evaluation. However, if there are several valid options, in order to determine the most interesting settings for the qualitative explorations, it seems reasonable to consider the corresponding classification results, i.e., to select those setting for further explorations which yield highest accuracies. The weights assigned by the classifier to the particular features in those settings are directly related to the observed accuracy. Yet, this is not necessarily the case if separate methods such as *Informations Gain* are applied to determine the most distinctive features for the given settings. These techniques might yield a different assessment of the contribution and importance of the particular features in a setting. So, there would be no direct connection between the features selected for analysis and the accuracy for the given setting. Yet, this connection should be retained, if accuracy is used as the criterion for selecting a particular setting for further explorations after all. Hence, using the weights assigned by the classifier as proposed above, seems a more consistent and easier interpretable approach.

Finally, using statistically more strict methods, such as the technique provided by the *Varb* family of programs (see Cedergren & Sankoff, 1974; Tagliamonte, 2012; Oliva & Serrano, 2013), might be of use here. However, e.g., strictly relying on statistical significance in qualitative explorations could be also misleading. On the one hand, for a data-driven approach as followed here, it might yield a vast amount of statistically significant effects which are not interpretable in any linguistically meaningful way though.[4]

[4]See, e.g., the discussion in Manning & Schütze (1999, p. 166), pointing to the fact that in the context of *collocation discovery*, only a small fraction of statistically significant observations seem to constitute reasonable collocation candidates. Thus in the qualitative analysis, the significance level is usually ignored, and only the ranking of the collocations based on the test statistic is

> "We believe statistical significance to be a secondary matter when describing variation, and especially when trying to explain it. So-called explanatory statistical tests may select a number of factors as co-occurring with some linguistic choice beyond chance, but this will not explain why such form is chosen in particular contexts nor the meanings that can be generated through its choice."
>
> (Oliva & Serrano, 2013, p. 64)

On the other hand, it might lead to hasty discarding effects determined as not significant (possibly due to data sparsity), and thus missing cases which can still be of interest and value for the research, and which could motivate further fruitful explorations in the given context.

Since in this chapter the focus is *not* on investigating the possible advantages and disadvantages of different evaluation techniques applicable in the given context, but rather on showing how the SLA research could generally benefit from a variationist approach to NLI, here we prefer the simple and accountable method described in the paragraph above. We show that it is well capable of discovering interesting interpretable *L1-transfer candidates* and thus of helping to advance the insight in SLA. Applying some more elaborated and strict techniques could be considered as a next step in the analysis though, i.e., for further validation and refinement of the findings.

10.2.5 Feature Grouping

Next, for our qualitative explorations, we used a more strict version of the feature grouping technique proposed in Section 8.3 and explored in detail in Section 9.3. In particular, due to conceptual considerations we made the clustering procedure more restrictive (see p. 113, point *intersection*):

1. Only variable groups where all of the variables show at least one common variant (i.e., the intersection of the variant sets is not the $\emptyset$) are kept as groups.

evaluated in the end.

2. Only those variants common to all variables in a group (i.e., the intersection of the variant sets) are kept as variants for that group. Any other variants are dropped.

While these restrictions clearly increase data sparsity, and also might miss some highly interesting generalizations, they reduce the number of cases which are linguistically not sound, i.e., it increases the probability that only linguistically meaningful combinations of variables and variants are generated by the grouping procedure. Thus, it makes the resulting groups of variables better interpretable, which has priority in this chapter.

10.2.6 Settings

Another important question is what results exactly to present based on the given features? There are many conceivable options, and we have to restrict them in some way to keep the procedure as comprehensible as possible. In this regard, the distinction between the different *levels of granularity* of linguistic variables suggested in Section 8.2.3 seems to be a reasonable basis. Given suitable linguistic variables, it allows for a systematic exploration, covering different conceptually interesting cases:

- **Evaluation setting 1 (ES1)**: Focus on particular variants at the most general level, i.e., exploring the variants in connection with most abstract and general variables (CGVs)
- **Evaluation setting 2 (ES2)**: Focus on particular individual variables and their variants at the most specific level (FGVs)
- **Evaluation setting 3 (ES3)**: Exploring the variables at some promising intermediate levels (IGVs) in between the two extremes ES1 and ES2.

In order to obtain the different evaluation settings, we first run the proposed label-informed feature grouping algorithm (see Section 8.3[5] and Section 10.2.5), and then select the settings of interest based on the output.

[5]Especially, cf. the discussion on the connection between the feature grouping technique and the level of granularity for linguistic variables on p. 116.

For ES1, we employ the setting corresponding to the *max. cut-off* in the dendrogram produced by the grouping algorithm, where all of the variables are automatically subsumed into a single cluster.

For ES3, we have to decide what exactly are the *most promising* settings. As already discussed in Section 10.2.4, the accuracy seems to be a reasonable criterion in the process of discovering settings which are potentially most interesting from the qualitative point of view. It is especially useful, if there are many conceptually equivalent alternatives, which is usually the case with IGVs – Applying the feature grouping technique, we obtain different sets of IGVs S at different cut-offs, with each S ($<$ max. cut-off) constituting a valid option for further explorations. Yet, by using the accuracy as selection criterion, we can simply pick the settings at cut-offs yielding the highest accuracies, and explore the corresponding set of IGVs in detail. This is the general procedure, we decided to follow in connection with the qualitative exploration of IGVs. The quantitative findings in Chapter 9 show that, generally, in order to discover the best settings for IGVs it is necessary to combine the variationist features with a surface model, restoring some of the properties glossed over by grouping – In particular, the settings [*s/c, +bvm*] performed always best in Section 9.3.4. Thus, here we follow this setup in the process of selecting the most promising settings for qualitative explorations using IGVs.

Finally, for ES2, we see two main options. The first option is to take the individual variables as they are provided to the grouping technique. The second option is to exploit the advantages of the grouping technique and to use the best performing settings for further explorations. In other words, to take the same setting as for ES3, namely, the setting at the best performing cut-off, but different from ES3, to consider the FGVs (clusters consisting of single variables) instead of IGVs (clusters consisting of at least two variables) contained in this settings. Here only those variables are kept as individual variables in the set which do not show quantitative advantages if combined into groups. This seems to constitute a reasonable feature reduction step, fostering the identification of interesting individual variables to explore. Thus, here we choose this second grouping-based

performance-driven option.[6]

In sum, ES2 (FGVs) and ES3 (IGVs) are based on the same setting, obtained at the *best performing cut-off*, while ES1 (CGVs) is based on the setting obtained at the *max. cut-off*. In ES2 and ES3 the variationist features are enhanced by a surface model, which in general seems necessary for identifying the best performing cut-offs. Whereas ES1 uses only variationist features – here adding a surface model does not seem to provide any practical or conceptual advantages for qualitative explorations.

10.3 Explorations

Following the conceptual considerations and implementing the decisions discussed in Section 10.2, we conduct a set of qualitative explorations utilizing selected variationist features, we consider among most suitable and promising in the given context. In particular, in Section 10.3.1 we explore *subject realization* patterns in general and the *subject pronouns* in particular, as complements; and in Section 10.3.2 we focus on exploring the choices in the context of the *nominal modification*, as adjuncts. In the following, we explore these phenomena and establish links between SLA and VS research. In particular, we exemplify the process of discovering potential new *L1-transfer candidates* using our approach, and exemplify how specific hypotheses about L1-transfer – suggested in the related research and new ones – can be tested and further explored using our method.

10.3.1 Subject (Pronoun) Realization

Introduction and Motivation The investigation of research questions related to the subject realization and, in particular the usage of pronouns as subjects has been targeted in various SLA and VS publications (see, e.g., Selinker & Lakshmanan, 1992; Gundel & Tarone, 1992; Wang, 2009; Domínguez, 2013; Shi, 2015; Brunner, 2015; Hacohen & Schaeffer, 2007; Giacalone Ramat, 2003; Oliva & Serrano,

[6]If it turns out that there are no FGVs in a setting (if there are only IGVs, i.e., clusters with at least two members), it is still possible to back-off to the first option. However, this was at no point necessary in our explorations.

2013; Tagliamonte, 2012). The research was mostly focused on specific phenomena, such as the acquisition of null and overt subjects or the *pro*-drop phenomenon in particular, concerned with the omission of pronouns in the subject position, etc.

In this section, we take a broad data-driven perspective on the question, integrating SLA and VS perspectives. First, we investigate the phenomenon of *subject realization* at an abstract level using POS. Second, we focus on *subject pronoun* realization in terms of individual surface forms selected by the learners. Furthermore, we show how particular hypotheses suggested in the previous research can be tested using our approach. Since the subject constitutes a part of the verb argument structure, from the functional point of view this investigation is concerned with usage patterns of *complements*.

Following the variationist perspective, here we consider the subject realization of an individual verb lemma as a linguistic *variable*, and the particular realizations as *variants*.[7]

For the given purposes, we utilize a version of variationist features, compiled based on the category realization and verb subcat features, systematically investigated from the quantitative perspective in Chapter 9. Essentially, we use a truncated version of the *SOX-POS-NoP* feature type introduced and evaluated in Section 9.3.4 (see also Table 9.2), which incorporates both, the required function and category realization information. By *truncated* we mean that only the subjects with the corresponding POS tags are kept as features, while the rest of the verb argument structure is dropped and not considered in this context. Where the information on particular surface forms is required, we consistently employ the lemmatizer provided by the *MATE* tools (see Section 10.2.2), and generate features incorporating the corresponding lemmas. We utilize all of the three setting as described in Section 10.2.6, namely ES1, ES2 and ES3 corresponding to the different granularity levels of the variables. In the following explanations we refer to various tables provided in *Appendix A*, which shows data relevant for our analysis, i.e., particular features, weights, groupings, etc.

[7]We provide the basic definition concerning *individual variables*, i.e., FGVs as targeted in the ES2 setting. The CGVs and IGVs targeted in ES1 and ES3 settings correspondingly, are obtained by different groupings of the FGVs (see Section 10.2.6).

Subject realization in terms of POS (*S-POS*) In this paragraph, we investigate the *subject realization* patterns in terms of POS following the different settings described in Section 10.2.6. The features are of the general form *s+POS* meaning the following: Subject – depicted by *s* – is realized by the given POS-tag – depicted by *POS*. Since we generate the features in a data-driven manner using standard NLP-tools always showing some error rates (tagging and parsing errors, etc.), there are usually also some corrupted features in the set, such as *s+"*, which are not interpretable in any linguistically meaningful way. In the following, we ignore these features in our analyses.[8]

ES1 The features representing the different variants extracted from the training data, and the corresponding weights for this setting are presented in Table A.1. The data provided in the table shows that at this most abstract variable level (CGVs) there are three variants which seem particularly indicative for L1 German, namely:

- *s+ex* (subject realized by the existential *there*)

  Assuming the collocation *there is* as the most common surface realization for this variant, its distinctive power could be explained by the high frequency of the equivalent *es gibt* in German (cf. MFGW, 2016). This particular effect was also discovered by Malmasi & Dras (2014d).

- *s+cd* (subject realized by a cardinal number)

  It is not obvious, why cardinal numbers as subjects might be indicative for L1 German in general. However, further analysis of the data suggests that this variant can be primarily related to the usage of the indefinite pronoun *one* as subject which has been tagged as cardinal number. Now, this finding becomes better interpretable – It can be related to the fairly frequent use of the equivalent *man* in German (cf. Tschirner & Jones, 2006; Brysbaert et al., 2011; MFGW, 2016), including translations from English to German (Johansson, 2007).

[8] Alternatively, as a further refinement step, such features could be eliminated by some pattern matching procedure, before running the classifiers.

- *s+dt* (subject realized by a determiner[9])

  The fact that this variant is indicative for L1 German could be attributed to the frequent use of the demonstrative pronouns such as *der/die/das*, *dieser/diese/dieses*, etc. in German (cf. Tschirner & Jones, 2006; Brysbaert et al., 2011; MFGW, 2016). However, it is just a vague assumption. At this point, it is not possible to make any stronger conclusions without further data.

In sum, the ES1 setting is useful for revealing some first interesting tendencies, and allows for first interpretation attempts in the context of L1-transfer. However, the employed features are very general, and a comprehensive and solid interpretation without any additional information is hardly possible. More data, first of all, more specific features are required for better interpretability of the outcome.

ES2 Since, investigating and interpreting all of the individual variables (FGVs) within a setting is hardly feasible, we can generally only focus on some selected cases, we consider worthwhile exploring here.

Based on the findings in ES1, we decided to test a specific hypothesis in connection with the variant *s+ex* using two particular variables, i.e., subject realization in connection with the two verb lemmas *be* and *exist*. The verb *be* plus the variant *s+ex* reflect the common surface realization *there is*[10], which was assumed in our interpretations in ES1. Yet, this variant as such does not directly include any information regarding the actual FGVs, i.e., particular verb lemmas, it was used with. Now, the German *geben* within the collocation *es gibt* corresponding to the English *there is*, has a meaning close to *exist* (Duden, 2003). So, our hypothesis is that learners with L1 German might use *there exists* instead, or besides the correct equivalent *there is*. This usage pattern might be indicative for L1 German constituting an interesting L1-transfer candidate. The weights for the different variants

[9] In the given tag set, the POS tag *dt* represents articles as well as some pronouns which do not have a separate tag, e.g., demonstratives. Since articles cannot function as subjects, we interpret the feature *s+dt* as representing pronouns.

[10] Note that since we use verb lemmas, this variant covers also all derived forms due to the conjugation of *be*, i.e., *there are*, *there was*, etc.

occurring in the context of the two variables under consideration are presented in Table A.2 and Table A.3. The data shows the following:

- *be_s+ex*

  There is a high weight for the variant *s+ex* in connection with *be* for L1 German: The weight is second highest across all 11 L1s (i.e, second highest across the 11 weights assigned by the separate classifiers to the given feature, showing that this features is in general important for distinguishing L1 German from the other L1s in the set), and it is the highest across all variants in connection with *be* (i.e., if *be* is used in L1 German essays, then it is most indicative in connection with the particular variant *s+ex*).

- *exist_s+ex*

  The data also shows a reasonably high weight for the variant *s+ex* in connection with *exist* for L1 German: The weight is third highest across all 11 L1s, and it is the highest across all variants in connection with *exist*.

On the one hand, it shows that the usage of *be* in connection with *s+ex* is highly indicative for L1 German. On the other hand, the results suggest that the usage of *exists* in connection with *s+ex* is indicative for L1 German as well. Thus, the findings support our hypothesis regarding the use of the collocation *there exists* as an alternative to the common form *there is* for L1 German, pointing to an interesting L1-transfer effect. However, one detail has still to be clarified: The actual surface realization of the variant *s+ex* is not accessible at this point, i.e., we cannot tell for sure that the first part of the two collocations of interest is the surface form *there* indeed. To investigate this point, more specific features are required. So, we target this issue later in connection with the *S-Pro-POS/L* features, providing the missing information (see p. 165).

ES3 The following groups (IGVs) are among the most indicative for L1 German (the groups are listed with the corresponding indicative variants):

- G1={*accept, define, govern, respect*}
  - *s+wp* (subject realized by a wh-pronoun)
- G2={*differentiate, fly, relate, suit*}
  - *s+wp* (subject realized by a wh-pronoun)
  - *s+wdt* (subject realized by a wh-determiner)
- G3={*arise, emerge, stem*}
  - *s+wdt* (subject realized by a wh-determiner)
- G4={*admire, avoid, fear, worry*}
  - *s+prp* (subject realized by a personal pronoun)

The weights for the different variants in connection with the groups G1, G2, G3 and G4 are shown in the Table A.4, Table A.5, Table A.6 and Table A.7 respectively.

Interestingly, the variants indicative in connection with the listed groups are all different from the variants, generally indicative for L1 German (see ES1, in particular Table A.1). This nicely highlights that the method supports the discovery of fine-grained differences between writers with different L1s in terms of the particular variants used in realizing different groups of verb lemmas as variables.

Particularly interesting are the groups G3 and G4, containing semantically related verbs, i.e., {*arise, emerge*} and {*avoid, fear, worry*} respectively. This shows that the technique is well capable of generating meaningful groups consisting of related variables, based on the variants usage.

The remaining question is, why exactly these groupings might be of interest in the context of L1-transfer? Apparently, there are some properties which make them indicative for L1 German. However, identifying these properties seems to require some more in depth analysis by SLA researchers. The groups generated by the technique, certainly seem to provide interesting "food for thought", providing data-driven inspiration for future theory-driven interpretation.

Testing an existing L1-transfer hypothesis about subject realization for L1 Chinese (E1+E3) In ES3 we showed that the proposed grouping technique is capable of generating meaningful groups of variables, providing potentially interesting candidates for further explorations on L1-transfer. In this paragraph, we sketch how the grouping technique can be used to advance the qualitative analysis in the context of the SLA research, by serving as a tool for testing and verifying existing hypotheses about L1-transfer. For this purpose, we employ the ES1 and ES3 logic, i.e., we use CGVs and IGVs.

In particular, we investigate the hypothesis by Wang (2009), suggesting that learners with *L1 Chinese* prefer *pronoun-subjects* over *noun-subjects*, leading to improper productions in Chinese-English translations:

> "Chinese people always hold the idea that human being and nature are mingled together, so Chinese people intend to make themselves as the start to narrate object things and are used to taking the pronoun as the subject. However, in the western philosophy, object is emphasized and it is believed that human being and nature are separated. So western people intend to express things from an object view and are used to taking non-pronoun such as things or abstract concept as the subject. The choice of pronoun-subject or non-pronoun-subject between Chinese and English will lead to negative transfer of mother tongue, which will make the translation of subject an improper one."
>
> (Wang, 2009, p. 139)

The findings of Wang (2009) based on translations by 81 students, support the hypothesis.

This specific hypothesis can be investigated by employing the features explored in this section. For this, we simply limit the set of variants to subjects realized by personal pronouns (*s+prp*) or nouns (*s+nn*), and employ the corresponding reduced feature set. Then we run *one-vs.-rest* classifiers with L1 Chinese vs. the western L1s in our set, namely French, German, Italian and Spanish following the ES1 and ES3 logic.

First, we explore the general usage pattern for *s+prp* and *s+nn* variants, detached from particular verb lemmas following the logic of ES1. This results in

a feature set consisting of just the two variants of interest, encoded by the relative frequency for each text. It turned out that the weights for both features are negative, showing that there do not seem to be any pattern indicative for L1 Chinese compared to the western L1s. Second, we follow the logic of ES3 and bring the actual verbs back into the equation. Here, all verb lemmas are grouped into five clusters (for the best performing cut-off). The findings are summarized in Table 10.1.

group id	# verb lemmas	pattern indicative for L1 Chinese	weight
G1	166	s+nn	< 0.01
G2	100	-	-
G3	312	-	-
G4	65	s+prp	0.73
G5	68	s+prp	0.19

Table 10.1: Usage pattern for L1 Chinese following the ES3 logic.

For G1 there is some very weak (a positive weight ≈ 0) indication for the variant *s+nn*, which essentially can be ignored. For the two groups G2 und G3 there is no indicative pattern for L1 Chinese, whereas for the two groups G4 and G5, there is a clear indicative preference for the variant *s+prp*. In sum, for most of the verbs in our data set, there is no indicative usage pattern for L1 Chinese compared to the western L1s. However, in connection with some particular verb groups, there is an indicative preference indeed, namely, for the variant *s+prp*, supporting the given hypothesis. Interestingly, the method does not simply support a known hypothesis, but it makes it possible to observe subsets of verbs for which the characteristics emerge. Studying what the 65 verbs grouped in the most indicative cluster have in common thus provides the opportunity for a more fine-grained qualitative analysis in SLA research.[11]

[11]The five groups are listed in Table A.8. Some of the members are not verbs, which was to expect due to the usual tagging error rates. Note that due to the RS3 (see Section 9.3.3), ensuring that only lemmas occurring repeatedly in the training data are considered, at least most of the idiosyncratic cases get filtered out here. In other words, only lemmas repeatedly mistagged as verbs in the training data, get into the feature set.

Subject *Pronoun* realization in terms of POS and lemma (*S-Pro-POS/L*) In the previous paragraph, we investigated the subject realization in terms of POS, which provides a interesting general perspective. We showed that the used abstract features can foster the discovery of interesting usage tendencies, and allow for first interpretations in the context of L1-transfer. However, the findings suggest that using some more specific features might further advance the qualitative insight. Thus, in this paragraph, we turn to qualitative explorations using a more specific feature set. The question is which features exactly to target in the given context?

The hypothesis by Wang (2009) tested above is concerned with the usage of pronouns vs. nouns as subject (see also Domínguez, 2013); other SLA work investigates the acquisition of null and overt subjects or the *pro-drop* phenomenon (see e.g., Gundel & Tarone, 1992; Selinker & Lakshmanan, 1992; Hacohen & Schaeffer, 2007); related VS work explores the particular choice of pronouns as subjects (Oliva & Serrano, 2013), or the connection between subject pronouns and complementizer realization (Tagliamonte, 2012), etc. Moreover, many of our interesting findings using *S-POS* features presented above (see p. 158), are related to the usage of pronouns as subjects, and further explorations employing more specific features would certainly help refining them. Thus, investigating the realization of *subject pronouns* in some more detail, seems a reasonable and promising choice in the given context. In addition, the pronouns constitute a closed class, which keeps the set of possible realizations (variants) comprehensible and the outcomes easier interpretable compared to open class features.

We use a modified version of the *S-POS* feature type, employed above. On the one hand, we utilize a reduced POS-tag set, limiting the variants to the generic pronouns, namely *prp* (personal pronoun), *prp$* (possessive pronoun), *wp* (wh-pronoun), *wp$* (possessive wh-pronoun), *ex* (existential *there*). In addition, we include the core determiner tag *dt*, which covers articles and some pronouns depending on the context – Since articles cannot function as subjects, we assume that subjects tagged *dt* are pronouns (e.g., demonstratives or quantifiers). Thus in sum, we employ subjects limited to six different POS-tags as described above. However, after the feature extraction process it turned out that there were no occurrences of *wp$* in the feature set. Thus, finally, we use the following set consisting of five variants:

- *s+prp*: subject realized by a personal pronoun
- *s+prp$*: subject realized by a possessive pronoun
- *s+wp*: subject realized by a wh-pronoun
- *s+ex*: subject realized by the existential *there*
- *s+dt*: subject realized by a determiner

ES1 The first interesting insight is that in our training data the five POS pronoun variants employed here, are realized by 210 different lemmas as shown in Table A.9. There are clearly some mistaggings such as *enjoy* as *dt*, or *human* as *prp*, etc. However, most of the lemmas seem to be misspellings of the correct forms, e.g., *onother* instead of *another*, *htey* instead of *they*, *theyr* instead of *their*, etc. This findings exemplify in an impressive way how much variance is contained in learner data, even with respect to a closed class such as pronouns. On the one hand, many (slightly) differing variants increase data sparsity issues. On the other hand, there might be some indicative patterns behind the misspellings pointing so potentially interesting effects.

The list of the most indicative variants for L1 German and the corresponding weights are presented in Table A.10. The outcomes further support our findings in S-POS.ES1 (see p. 159) and provide some new interesting insight:

- *s+prp+one*

  In particular, the highly indicative variant *s+prp+one* supports our hypothesis regarding an L1-transfer effect, related to the frequent use of the equivalent *man* in original German texts as well as in the translations from English to German (cf. Tschirner & Jones, 2006; Brysbaert et al., 2011; MFGW, 2016; Johansson, 2007).

- *s+dt+this*, *s+dt+both*, *s+dt+neither* and *s+dt+the*

  In connection with S-POS, it also turned out that the variant *s+dt* was highly indicative for L1 German. Now, the given indicative variants

allow for further interpretations on this phenomenon. In particular, the high distinctive power of the variant *s+dt+this* provides a further important piece of information. Our hypothesis is that this might be related to the very frequent collocation *das ist* in German (cf. GBN, 2016), corresponding to *this is* in English, applicable in many different contexts. We further investigate this issue below in ES2. Furthermore, the indicative variants *s+dt+both* and *s+dt+neither* are both pronouns, corresponding to the frequent equivalents *beide* and *kein* in German (Brysbaert et al., 2011; Tschirner & Jones, 2006; MFGW, 2016), pointing to other interesting L1-transfer candidates. Finally, since articles cannon function as subjects, the variant *s+dt+the* seems to pose a tagging or parsing error. Further explorations show that this phenomenon can be often related to a misspelling of the personal pronoun *they* as *the*, leading to its mistagging as *dt* instead of *prp*, which turns out to be indicative for L1 German.

- *s+ex+threre*

  The indicative variant *s+ex+threre* points to the discussion in S-POS.ES2 (see p. 160) regarding the actual surface realization of the variant *s+ex*. We explore this issue below in ES2.

- *s+prp+zou*

  A special case is the indicative variant *s+prp+zou*. This does not seem to be related to the L1-transfer as such, but rather to a sort of *"cultural transfer"*. One of the main differences between the German and English keyboard layouts is the position of the keys *y* and *z* – these two are simply swapped. Thus, it is fully comprehensible that L1 German learners would misspell *you* as *zou*, especially under a possible time pressure, emerging in the context of test such as TOEFL11, which is the source of the essays in our corpus.

ES2 Here, we further explore some of the findings presented above at the FGV level, in particular we target the variants *s+ex+there*, *s+dt+this* and *s+pro+one*:

- *be/exist_s+ex+there*

  As pointed out in ES1, we further investigate the variant *s+ex* by exploring its actual surface realization (see the discussion in S-POS.ES2, p. 160). Our results in S-POS.ES1 (see p. 159) suggest that the variant *s+ex* is highly indicative for L1 German. Yet, as already discussed in ES1 and shown in Table A.9, this variant is realized by various lemmas in our training data. Furthermore, the data in the Table A.10 suggest that the variant *s+ex+threre*, which constitutes a misspelled version of *s+ex+there* seems to be an indicative pattern for L1 German in general. So, the question is what particular realizations of *s+ex* make this variant indicative for L1 German? In order to further investigate this question, we inspect the weights assigned to the different variants in connection with the verb lemmas (variables) *be* and *exist*, following our exploration in S-POS.ES2. Table A.11 shows the most indicative variants realized with *be*, and Table A.12 provides the corresponding information for *exist*. It turns out that the most indicative realization of the variant *s+ex* is indeed *s+ex+there* for both of the verb lemmas. On the one hand, it supports our assumptions regarding the collocation *there is* as equivalent to the frequent German *es gibt* in S-POS.ES1. On the other hand, it further supports our findings in S-POS.ES2 regarding the collocation *there exists* as another possible equivalent to the German *es gibt* pointing to an interesting L1-transfer candidate – In particular, the outcome clarifies that the actual realization indicative for L1 German is *there* indeed, and not any other (possibly idiosyncratic) realization of the POS-tag *ex*, occurring in the training data (see Table A.9). Finally, The misspelled variant *s+ex+threre* seems to play some role in connection with the verb lemma *be* (second most indicative variant for L1 German), while it is not part of the realization frame of the verb lemma *exist*. In general, this misspelling is one out of many features which seem difficult to meaningfully interpret in terms of L1-transfer despite being indicative. This exemplifies that the technique is only capable of providing lists of *potentially* interesting

L1-transfer candidates, which then have to be analysed and evaluated by researchers.

- *be_s+dt+this*

  We further investigate our hypothesis presented in ES1 that the high distinctive power of the variant *s+dt+this* might be related to the common collocation *das ist* in German, corresponding to *this is* in English, applicable in many different contexts. Table A.11 indeed shows that *s+dt+this* is highly indicative for L1 German in connection with the verb lemma *be*, which support our hypothesis and points to another interesting L1-transfer candidate.

- *X_s+prp+one*

  The findings in ES1 further confirmed our assumptions in S-POS.ES1 (see p. 159) on the indicative use of the indefinite pronoun *one* for L1 German – The variant *s+pro+one* shows the highest weight across all variants indicative for L1 German in ES1 (see Table A.10). Interestingly, it turned out that this variant occurs in connection with only a small set X of variables in our training data, namely, with only nine verb lemmas. This makes it feasible to investigate the full usage pattern for this particular variant across all relevant variables, what we do here. The verb lemmas (variables) and the corresponding weights for the particular variant *s+prp+one* are presented in Table A.13. There are two particularly interesting findings: First, this variant is indicative for L1 German with six our of nine verb lemmas, thus with most of the variables. Second, the six verb lemmas of interest are the following: *can*, *could*, *have*, *might*, *must*, *should* – All of them express a modal meaning. Thus, it seems that L1 German learners tend to prefer impersonal constructions in realizing modal meanings. Since the corresponding collocations in German, namely, *man kann*, *man könnte*, *man hat (zu)*, *man dürfte*, *man muss*, *man sollte* are very usual (cf. GBN, 2016), it seems to be an interesting L1-transfer candidate.

ES3 Here, we explore whether our qualitative findings presented and discussed above could be further enhanced by some IGVs. In that regard the following groups seem interesting (the groups are listed with the corresponding indicative variants):

- G1={*allow, imply, motivate, prevent*}
  - *s+dt+this*
- G2={*count, increase*}
  - *s+dt+this*
  - *s+prp+i*
  - *s+prp+they*
- G3={*give, happen*}
  - *s+ex+there*
  - *s+dt+that*
  - *s+wp+who*

Table A.14 and Table A.15 show the variants and the weights for the groups G1 and G2 respectively. The variant *s+dt+this* is indicative for both groups, showing that there are some potentially interesting usage patterns in terms of L1-transfer beyond the collocation *this is* as discussed in ES2. The Table A.16 provides the corresponding information for the group G3. It turns out that the variant *s+ex+there* is the second most indicative across all possible variants for G3. Thus, this verb group poses a potentially interesting candidate for further explorations on the given variant, beyond the verb lemmas *be* and *exist* explored in detail above (see ES2).

These findings suggest that our approach can also support explorations in the context of the *construction grammar*, pointing to specific *constructions* (i.e., form-meaning mappings which are conventionalized in the speech community), common for learners with particular L1s in their L2 productions, possibly due to L1-transfer effects (cf. Ellis, 2013).

10.3.2 Nominal Modification

Introduction and Motivation In Section 10.3.1, we explored features realizing complements, namely, realization of subjects as part of the verb argument structure. In this section, we turn to the opposite part, and explore features realizing *adjuncts*, thus further broadening the view. The question is whether we could also capture some interesting patterns in the variation of structurally non-mandatory units? In particular, we explore the *nominal modifiers* as adjuncts.[12] This also means that we move the research focus from the verb to the noun domain, which also contributes to broadening the view on our approach. After investigating some general usage tendencies utilizing POS, we explore a particular hypothesis regarding the position of the modifiers with respect to the head nouns. Our agenda fits well into the SLA context, with a range of contributions concerned with nominal modification (e.g., Brunner, 2015; Vyatkina et al., 2015; Hirschmann et al., 2013; Yang, 2014; Kanehira, 2003) and general word order regularities (cf., e.g., Gass & Selinker, 1983; Odlin, 1989; Choroleeva, 2009; Domínguez, 2013; Shi, 2015).

Following the variationist perspective, here we consider the particular modifier realization in connection with an individual head noun lemma as a linguistic *variable*, and the particular realizations as *variants*.[13]

Here, we focus on the setting ES1 as described in Section 10.2.6. For the given purposes, we utilize variationist features, which are conceptually different, but technically similar to those explored in Section 10.3.1. The main difference is in the POS tag of the units constituting the variables, i.e., nouns instead of verbs,

[12] In fact, some nominal modifiers have properties of complements. However, the cases where a modifier is structurally mandatory (e.g., *They are denizens of the forest*) are rather rare, and in addition, these are not as clearly differentiated syntactically as in the clause structure (Huddleston & Pullum, 2002). Thus, for simplicity, we consider all modifiers as adjuncts here.

[13] This is again the basic definition, reflected by FGVs targeted in the ES2 setting. The definitions for the other granularity levels can be inferred from this basic one (see Section 8.2.3 and Section 10.3.1). Note that the specification of what poses a variable and what poses a variant seems controversial in this context. In particular, since the heads select their complements, whereas adjuncts seem to select their heads (see the discussion in Pollard & Sag, 1994), it might be more appropriate to treat the individual modifiers as variables and the corresponding head nouns they occur with as variants. However, since we are interested in the patterns of *modifier variation* rather than noun variation, we flip the perspective, and use the definition proposed above, according to our research interest. Since it allows to investigate a particular relevant research question, we consider our definition as appropriate in the given context.

and in the functions under consideration constituting the variants, i.e., nominal modifiers (*nmod*) instead of subjects. In addition to the two core differences, there are two further modifications:

1. *Head noun position*: We encode the position of the head noun relative to the modifier, in order to be able to distinguish between pre- and post-head modifiers, or more precisely, between *pre-nominal* and *post-nominal* modifiers (see Huddleston & Pullum, 2002).

2. *Level of feature specificity*:

   (a) For our qualitative explorations in Section 10.3.1, we used a truncated version of the *SOX-POS-NoP* feature type (employing only subjects), where the core POS categories (verbs, nouns, adjectives and adverbs) were represented by a single tag. This reduction was introduced in connection with the *SOX-POS-NoP* feature type, in order to limit data sparsity. Here the situation is different. On the one hand, nouns are generally more frequent than verbs, which is expected to reduce negative data sparsity effects compared to the verb subcat features. On the other hand, in this context the advantage of employing a more specific tag set might outweigh the disadvantages resulting from potential data sparsity issues: E.g., it might be of high interest whether an adjective as modifier is used in the positive, comparative, or superlative form, or whether a present of past participle is employed for modification, etc. Thus, here we employ the full *PennTreebank* tagset (Santorini, 1990)[14] in the feature generation process for our explorations in ES1a.

   (b) In ES1b, we turn to the opposite, and use maximally abstract features, where each modifier is represented simply by the token *nmod*. In other words, in ES1b we only encode the position of any modifier relative to the head noun.

The required features can be easily generated employing previously used tools (see Section 10.2.2). As in Section 10.3.1, if the information on particular surface

[14] `https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`

forms is required in a setting, we consistently employ the lemmatizer provided by the *MATE* tools (see Section 10.2.2), and generate features incorporating the corresponding lemmas. In the following explanations, we refer to different tables provided in *Appendix B*, showing data relevant for our analyses.

ES1a All variants occurring in our training data, and the corresponding weights for the different L1s are presented in Table B.1. Table B.2 poses a reduced version of Table B.1, showing only those variants which are most indicative for L1 German.

- Participles
  - *vbn+N* (pre-nominal past participle)

    Past participles functioning as pre-nominal modifiers constitute a common pattern in German (cf., e.g., Eisenberg, 1994). Depending on the genre and register, the corresponding constructions can vary in terms of frequency and complexity. In fact, participles as modifiers belong to the group of forms which are mainly responsible for the complexity of the sentence structure in German (Weber, 1971). It seems reasonable that L1 German learners would transfer such constructions to a L2. Moreover, *vbn+N* shows the highest weight across all modifier variants indicative for L1 German. In sum, it seems to pose a promising L1-transfer candidate.

- Adverbs
  - *N+rb* (post-nominal adverb)
  - *rb+N* (pre-nominal adverb)
  - *wrb+N* (pre-nominal wh-adverb)

    The usage of adverbs as nominal modifiers seem to play some role in distinguishing L1 German from the other 10 L1s in the set, pointing to a potential L1-transfer effect. In fact, adverbs can be used as post-nominal modifiers in German, but the options here are rather limited (Eisenberg, 1994). Thus, at this point it is

not clear how to interpret the indicative patterns presented above. Further explorations are required to clarify this issue.

- Hyphens
  - *N+hyph* (post-nominal hyphen)

    Interestingly, this variant shows the second highest weight across all modifier variants indicative for L1 German. Since noun compounding is one of the most productive grammatical processes in German (Hieble, 1957; Voyles, 1967), and the usage of hyphens is closely related to that phenomenon, this finding seems to point to another interesting potential L1-transfer effect.

ES1b As mentioned above, in this setting, we use a heavily reduced, abstract version of the features employed in ES1a, where the only encoded information is the position of a modifier relative to its head noun (see p. 172). In other words, there are only two variants as features, namely:

- *H+NMOD*
- *NMOD+H*

We use these features to test a specific hypothesis concerting L1 German vs. L1 French productions in English. In particular, in French the typical word order is *modified-modifier*, whereas in German the opposite holds, i.e., *modifier-modified* is usual (Bally, 1944; Greenberg, 1971, p. 52). Our hypothesis is that this tendency might be reflected in L2 English.

Running a binary classifier for L1 German (positive class) vs. L1 French based on the T11 data, yielded the feature weights as presented in Table 10.2.

Variant	GER
H+NMOD	-0.43682
NMOD+H	0.35743

Table 10.2: Weights assigned by a binary L1 German vs. L1 French classifier to the positional NMOD features.

In general, the usage of pre-nominal modifiers (*NMOD+H*) seems to be indicative for L1 German showing a reasonably high positive weight. Whereas the usage of post-nominal modifiers (*H+NMOD*) contributes to a classifier decision towards L1 French showing a comparable negative weight. This usage tendency supports our hypothesis, pointing to another interesting L1-transfer candidate. In fact, some findings in the related SLA research point into the same direction – In particular, the placement of adjectives by francophone learners of English seem to show a notable tendency towards the post-nominal position, which is rather unusual for native English productions (cf. Choroleeva, 2009).

10.4 Conclusions

Complementing the quantitative explorations in Chapter 9, in this chapter, we focused on the qualitative aspect of our variationist approach to NLI. For that, we proposed and discussed an easily interpretable method utilizing logistic regression weights assigned to linguistic variables at different levels of granularity as introduced in Section 8.2.3. On the one hand, we exemplified how new L1-transfer candidates can be discovered in a data-driven way based on the variationist perspective. In particular, we used a top-down procedure, showing how to obtain first promising findings, and how using different settings, these findings can be gradually refined and explored in more detail, leading to valuable new insights. On the other hand, we showed how the method can be employed to test theory-driven hypotheses about L1-transfer, supporting the validation of existing findings and new hypotheses. The method seems to provide a fruitful integration of ML techniques and a variationist perspective, well capable of advancing the SLA research.

Now, lets turn to the provocative question in the title of this thesis, suggesting if there is real choice in the production of non-native output, or if the choice is predetermined by the L1. Consider, e.g., the variants *there is* vs. *there exists*, both used by L1 German learners in L2 English apparently as equivalent to the German *es gibt*, whereas *there is* is a regular realization and *there exists* is rather unusual. Thus, one could argue that the learners, in a way, even have an *"greater choice"* in their non-native voice compared to the natives: The learners usually show regular plus various irregular realizations, thus in fact, they seem to have an extended set

of options compared to the natives, generally realizing regular forms. Identifying the preferences in the regular realizations as well as the indicative irregular cases, remains a highly interesting and thrilling research direction. We hope that our method will pose a useful tool for discovering new, valuable findings on that in the future work.

Connecting the findings and the research questions Regarding the particular five research questions in the focus of this thesis (see Section 1.3), the discussion in this chapter contributes to three of them as follows:

1. [VARIATIONIST-PERSPECTIVE]: The whole chapter was essentially dedicated to the clarification of the second part of this research question, i.e., to the exploration of a variationist approach to NLI from the qualitative point of view, and its general value for the SLA research. Utilizing suitable linguistic features and investigating them in detail under a variationist perspective, led to instructive and valuable qualitative findings. On the one hand, it fostered the discovery of new interesting L1-transfer candidates in a data-driven way. On the other hand, it allowed for testing and verifying theory-driven hypotheses on L1-transfer. Combining both directions is essential for advancing the insights in SLA.

2. [LINGUISTIC-FEATURES]: In this chapter, we explored new linguistic features following the variationist perspective, namely *subject (pronoun) realization* and *nominal modification*, focusing on the qualitative aspect of NLI. Their evaluation contributed new valuable insight to the SLA research.

3. [GENERAL-STRUCTURES]: As discussed in Chapter 8, it is possible to abstract over the individual linguistic features by using grouping techniques. The findings in this chapter show how such abstraction can help advancing the insights into the general underlying structures reflected in NLI. Namely, the abstract CGVs can pose an important first step for the qualitative analysis, pointing to general interesting cases, which then can be refined in further explorations using more fine-grained units such as FGVs and IGVs. In general, combining abstract linguistic information with surface properties yields best interpretable qualitative findings.

Chapter 11

Advantages and Limitations of the Variationist Approach

In the current part of this thesis, so far we proposed a variationist approach to NLI, and described as well as evaluated its possible implementation in both, quantitative and qualitative regards. In this last chapter, we conclude by summarizing some of the advantages and limitations emerged from our study.

Advantages Our approach shows several valuable advantages in conceptual, quantitative as well as qualitative regards.

- *Conceptual advantages*: The proposed approach enables a specific view on language units at different linguistic levels. In particular, it allows for considering and exploring a set of conceptually, structurally or contextually related units in direct connection to each other. This is realized by utilizing the notion of a linguistic variable and the variants realizing it. If the set of variants realizing a particular variable cannot be determined in advance, or if it seems beneficial not to restrict this set a priori, the theory also allows for a more relaxed view, by relating each variant to the set of all potential variants. Moreover, it is possible to infer new, more general linguistic variables, based on the original set of variables. For this, we proposed a flexible method using ML techniques, in particular clustering.

- *Quantitative advantages*: Given variationist feature encodings, the classifier can figure out most indicative variant distributions for different L1s, which can lead to advantages in terms of classification performance.

- *Qualitative advantages*: A variationist perspective allow to concentrate on the proportions of related variants in the data, which can help discovering new interesting usage patterns for learners with different L1s in a data-driven way. Moreover, such a view on the features allows for formulating and testing various theory-driven hypotheses about L1-transfer, assuming different variant preferences for particular linguistic variables across learners with different L1s. It is also possible to combine the data- and theory-driven procedures, i.e., one could identify interesting variant preferences in a data-driven way first, and then use this findings to formulate and validate new theory-driven hypotheses about L1-transfer.

Limitations However, the approach shows also certain limitations.

- *Conceptual limitations*: Here, we utilized a relaxed definition of a linguistic variable, not imposing any semantic constraints on the variants, such as the requirement of the variants' meaning equivalence, usually imposed in the VS research. The proposed method does not provide any means to ensure such properties. While we argue that always keeping the traditional requirement of meaning equivalence is not necessarily beneficial in any context (in fact, sometimes it also might be an unnecessary limitation), we believe that extending the approach by some components for semantic analysis, could be an interesting direction worth pursuing in terms of the future work though. It would provide a new set of potentially interesting options.

- *Quantitative limitations*: There are some limitations in connection with the proposed label-informed feature grouping technique aimed at automatically generating linguistic variables at more abstract levels. From the quantitative point of view, the technique usually does not show any noteworthy advantages if applied without any further steps. We assume the following: While feature grouping as proposed here, is capable of reducing data sparsity by

considering groups of related variables instead of individual, possibly rare ones, it apparently also hides some specific traits of the individual variables, which can be indicative and thus useful for classification. We showed that adding suitable surface-based models to the system, capable of restoring the cues glossed over by grouping, is indeed helpful, leading to an improved performance. Another related issue: The more specific the features become, the less seem the advantages of the grouping, even if its supported by corresponding surface models. For the classifier, it is apparently more suitable to have the access to the individual specific features, and to weight them separately. In this way it can discover some highly indicative cues reflected by certain combinations of individual variables and variants which get lost by considering groups of variables an their variants. In sum, grouping proves most advantageous if it gets applied to some more general and abstract linguistic features, and at the same time gets supported by appropriate surface models. In addition, the performance gains of around 2% in terms of accuracy, we were able to obtain by grouping variables, are not as high as expected. However, we showed that this could be attributed to the rather limited size of the T11 data used in our explorations.

- *Limitations of the qualitative analysis*: From the qualitative point of view, there is an issue regarding the interpretability of the more abstract variables, induced by the proposed automatic feature grouping procedure. While for some groups, the contained variables seem to be related, e.g., they show some common properties, most of the groups are rather difficult to interpret. However, this does not mean that a meaningful interpretation is not possible though. It might require some more in depth analysis for identifying the particular properties making those groups indicative for particular L1s. Another issue is that the indicativeness of the features in the employed settings, always means that these are indicative in relation to the other L1s in the particular given set. Thus, depending on the data, the outcomes might differ. However, since we employed a data set including a reasonable number of very different L1s, we believe that our findings are conclusive though. Further, in our qualitative analysis we consciously avoided any strong con-

clusions about L1-transfer, speaking about L1-transfer *candidates* instead. There are three main reasons for this: First, while the used data set, which was compiled specifically for the task of NLI, namely TOEFL11 (T11), seems generally to be of high quality, controlling for relevant parameters such as the topic bias, we believe that it is not sufficiently large for any strong claims. Various features show rather low frequencies, especially if they get more specific. Using, e.g., the TOEFL-Big corpus (Tetreault et al., 2012), might help solving this issue. We hope that it will be made publicly available in the future. Second, there seem to be various tagging and parsing errors resulting from the properties of the learner data, which seems generally challenging to process for standard tools. Using tools which are better tuned to this specific language variety, would certainly help reducing these issues. Third, we believe that any findings should be verified on a range of appropriate data sets, using different methods, before making any strong conclusions. In practice, it is seems rather hard to meet all of these requirements. Nevertheless, we believe, this is necessary in order to ensure high quality of the findings, and thus should be of high priority in the future work.

Part IV

Advancing Performance

Chapter 12

Introduction

In this part of the thesis, we turn back to the quantitative perspective, and target the question how we could further improve the classification performance of a NLI system?

First, we incorporate a wide range of features utilized in our work, including the variationist features explored in the previous part. In particular, we are interested in the question what can variationist features as explored here, contribute to the overall performance of a NLI system. Can they provide a quantitative edge?

Second, incorporating a wide range of features raises the question how the different sources of information are best combined. The most simple solution is to put all available features into a single vector. However, Tetreault et al. (2012) showed that the performance can be increased by using a meta-classifier, in particular, a *probability-based ensemble* (see also Cimino et al., 2013, and Section 6.6.2). In addition, our findings in Chapter 6 suggest that employing some model subsets might improve the classification accuracy over incorporating all models into the system. But which models are worth integrating into such a meta-classifier? On the one hand, some of the models may be redundant despite performing well individually. On the other hand, some models may improve the ensemble despite performing relatively poorly by itself. We explore this issue by implementing a linear ensemble optimization algorithm performing model selection. Moreover, we propose how the ensembles can be further tuned.

In Chapter 13, we discuss our ensemble approach, before turning to its ap-

plication and evaluation in Chapter 14. The presented work mainly reports our findings in Bykh & Meurers (2014) and Bykh & Meurers (2016), extending them by some additional evaluations and discussions.

Chapter 13

Ensemble Optimization and Tuning Approach

In this chapter, we present and discuss our ensemble optimization approach in detail. In particular, we explore using a simple linear ensemble optimization algorithm performing model selection, and show how such classifiers can be further tuned.

13.1 Generating Ensembles

As pointed out in Chapter 12, to combine the individual models, we employ a probability-based ensemble approach, following Tetreault et al. (2012), which we already successfully used in Chapter 6.

The meta-classifier combines the probability distributions provided by the individual classifier for each of the incorporated models as features. Each probability distribution yielded by a individual classifier represents the probability estimates for the different classes (L1s) in the training set. To obtain the ensemble training files, we performed 10-fold cross-validation for each model on the corresponding training set, and took the probability estimate distributions. For testing, we took the probability estimate distribution yielded by each individual model trained on the corresponding training set and tested on the T11 *test* set. To obtain the probability estimates for the individual models we used *LIBLINEAR* (Fan

et al., 2008). The ensembles were trained and tested using LIBSVM with an RBF kernel (Chang & Lin, 2011), which outperformed LIBLINEAR for this purpose.

13.2 Ensemble Optimization

The growing range of features used for NLI raises the question of how to perform model selection. Even when analysing a single complex feature type such as *CFGR* in depth (see Chapter 9.2), we already have to determine which of the low-performing models to keep in an ensemble. We approach the question with a simple linear ensemble optimization algorithm performing model selection.

Algorithm 1 Ensemble Optimization / Ensemble Model Selection

```
M_a ← {m_1, ..., m_n}          ▷ overall ensemble, i.e., all ensemble models
M_b ← ∅                        ▷ current best ensemble
while M_a ≠ ∅ do               ▷ iterate until M_a is empty
    m_b ← MAX(M_a)             ▷ get the best model m_b out of M_a
    M_t ← M_b ∪ {m_b}          ▷ join the previous best ensemble M_b and {m_b}
    if ACC(M_t) > ACC(M_b) then ▷ check if the new ensemble M_t is better than M_b
        M_b ← M_t              ▷ if accuracy improves, store M_t in M_b
    end if
    REMOVE(m_b, M_a)           ▷ remove m_b from M_a
end while
```

In each iteration step, the optimization algorithm as defined in Algorithm 1 retrieves the current best single model m_b out of the model set M_a (which is initialized with the overall model set for a particular setting), joins it with the previous best performing ensemble M_b (which is initialized to $\emptyset$), and then compares the accuracy of that new ensemble to the accuracy of the previous best ensemble. It retains the new ensemble as the best ensemble if the accuracy improves, or keeps the previous best ensemble as best ensemble otherwise. In Algorithm 1, we describe only the gist of the optimization, omitting some details to keep it transparent. Some ambiguities have to be resolved. If there are several models in M_a yielding the same accuracy, one has to decide, which of them to pick as the next m_b. We resolve that issue by always picking the model with the least number of features. When several models yield the same accuracy and have the same number of features, we resort to alphabetical order. The optimization is always carried out

using 10-fold cross-validation results on the training data (to obtain the accuracy ranking on M_a and to perform each optimization step). The *test* set is not part of the optimization at any point. Only after optimization, the resulting ensemble is applied to the *test* set and we report the corresponding accuracies.

In the following explanations and tables, the shorthand +*opt* means that the optimization procedure described in this section is applied in a given setting. Correspondingly, the shorthand -*opt* means that it was not applied in a given setting.

13.3 Ensemble Tuning

In order to further tune the ensemble, we explore the following idea: We generate a *single ensemble model* m_{n+1} based on *all* of the features used in a particular setting, i.e., all the features incorporated by the models $m_1 \ldots m_n$. Then we include that m_{n+1} model in the M_a ensemble as just another model, and use that new M_a^{+1} ensemble either directly or as basis for the optimization. Since m_{n+1} incorporates all of the features of interest for a particular setting, it is expected to yield more reliable probability estimates than the other individual ensemble models in M_a^{+1}, each covering only a subset of that feature set. Incorporating such an m_{n+1} into the ensemble may stabilize the resulting system, i.e., the machine learning algorithms may learn to rely on m_{n+1} in settings where the rest of the included models $m_1 \ldots m_n$ show a rather poor individual performance and are of limited use. In the following explanations and tables, we refer to the model m_{n+1} as [*all*] and to the M_a^{+1} ensemble as +*all*. Correspondingly, -*all* means that we use just M_a, i.e., an ensemble without the [*all*] model.

For building the m_{n+1} model included in the M_a^{+1} ensemble there are two options. We can build it on the basis of the probabilities of the models or on the union of the original feature values of those models. In the former case, the final ensemble model essentially is a meta-meta-classifier. For the settings integrating the same type of feature representations (cf. results in Tables 14.1 and 14.3), we use the original feature values merged into a single vector to build m_{n+1}. For the settings integrating different feature types (cf. results in Table 14.5), we use the probability estimates from the models $m_1 \ldots m_n$ to build m_{n+1}.

13.4 Conclusions

In this chapter, following the findings by Tetreault et al. (2012) and based on our results in Chapter 6, we proposed an probability-based ensemble classification approach, utilizing ensemble optimization and tuning techniques. We believe that it can be helpful in dealing with increasing numbers of features, employed for NLI. On the one hand, it could help reducing the complexity of the systems by suggesting an appropriate subset of the models. On the other hand, it could lead to an improved performance by stabilizing the ensemble and reducing the noise in the feature space. We implement and evaluate the method in Chapter 14.

Connecting the findings and the research questions Regarding the particular five research questions in the focus of this thesis (see Section 1.3), the discussion in this chapter contributes to one of them as follows:

1. [MODEL-OPTIMIZATION]: In this chapter, we focused on presenting an approach which in our opinion is capable of optimizing large NLI systems incorporating a broad range of features. For this, we utilize ensembles (meta-classifier), combining probability-estimates for the different classes in the data, yielded by individual classifiers representing different feature types. We propose a method for optimizing such ensembles by utilizing a linear model selection algorithm. Moreover, we suggested a tuning procedure which is based on including all of the available features in an abstract, condensed form into the overall ensemble. This is expected to stabilize the classifier, i.e., to make it more robust. While the tuning could stabilize the ensemble leading to a better performance, the optimization algorithm might further improve the classifier by reducing the noise or redundancy. The interplay of both components could help optimizing complex NLI systems, and further advancing the performance. We evaluate the method in detail in Chapter 14.

Chapter 14

Advancing Performance Using Ensembles

In this chapter we implement and evaluate our ensemble approach, presented and discussed in Chapter 13. We show that applying the method to the proposed features yields state-of-the-art results, outperforming previous systems.

14.1 Data

The research in this chapter employs two data sets used in Chapter 6 (see Section 6.3) in the context of the First NLI Shared Task (see Section 3.2), namely T11 only for single-corpus, and NT11 plus T11 for cross-corpus evaluations. The data splits follow the settings used for the *closed* and *open-1* tasks in the context of the First NLI Shared Task, namely:

- *Single-corpus*: T11 *train* $\cup$ *dev* sets for training, and T11 *test* set for testing
- *Cross-corpus*: NT11 set for training, and T11 *test* set for testing

14.2 Features

We use a range of features, explored in the previous chapters of this thesis, and investigate the effect of combining them into different ensembles. In particular, we

employ a range of features introduced and discussed in Chapter 5 (different types of recurring n-grams), Chapter 6 (several linguistically-motivated feature types based on dependency, constituency, morphology, linguistic complexity, etc.) and Chapter 9 (syntactic variationist features).

Related research, first of all, the results of the First NLI Shared Task (see Section 3.2) suggest that n-gram features are best performing for NLI. Our findings in Chapter 6 support this conclusion. In addition, our results show that using some linguistically-motivated features, which in general perform worse than n-grams, on top of n-grams, can further improve the performance. Thus, in order to obtain high performing NLI systems, it seems reasonable to use n-grams as basis, and then to extend it by various linguistic features.

N-grams as basis We form our basis model by utilizing 40 different n-gram models. Adapting the n-gram approach we presented in Chapter 5, we use all *recurring n-grams* with $1 \leq n \leq 10$[1] at four different levels of representation. First, we include the word-based (word/W), open-class POS-based (OCPOS/OP) and POS-based (POS/P) n-grams from our previous work. In addition, we utilize a version of recurring n-grams, not used in our explorations so far, but which turned out to be useful in related research: lemma-based (lemma/L) n-grams. These were used among other features by the winning system of the First NLI Shared Task (Jarvis et al., 2013, see Section 3.2). We employ binary feature encoding for all n-gram types.

Linguistically-motivated features as extension We use a range of features explored in the previous chapters as extension to our n-gram basis. In particular, we employ the variationist features explored in Chapter 9, i.e., the *CFGR* feature type as well as two version of the *verb subcat* features – the basic and most abstract *SOX-NoP*, as well as the best performing and most specific feature type in our set

[1]Our results in Chapter 5 suggest that n-grams of $n > 5$ seem to be of limited use. However in this chapter, where we focus on model combination and selection, we decided to include longer n-grams up to $n = 10$ as well. On the one hand, the data set used here is bigger compared to the experiments in Chapter 5, which might be an advantage for the performance of longer n-grams. On the other hand, it might well be that some selected models based on higher n are still useful, if combined with some other models.

SOX-POS-NoP. Moreover, we employ the 16 core feature types as introduced in Chapter 6 (see Table 6.3).[2]

14.3 Tools

We do not employ any new tools here, i.e., we only utilize the tools introduced in the previous chapters in connection with the features and techniques which are relevant in this context (see Section 14.2 and Chapter 13).[3]

For the single new feature type used in this chapter, i.e., the *recurring lemma-based n-grams*, we employ the same lemmatizer as in Section 9.3.2, namely, the lemmatizer provided by the *MATE* tools (Björkelund et al., 2010).

14.4 Results

In this section, we implement and evaluate our ensemble approach presented in Chapter 13. We provide single-corpus (*sc*) and cross-corpus (*cc*) results for different ensembles, where *+/- opt* states whether ensemble optimization was performed, and *+/- all* whether tuning was employed as described in Chapter 13. Concretely, *(-opt, -all)* means that the ensemble M_a was used without any optimization or tuning, and correspondingly *(+opt, +all)* means that the optimized and tuned version of M_a (i.e., the optimized version of the ensemble M_a^{+1}) was employed. In the remaining two cases *(+opt, -all)* and *(-opt, +all)* either optimization or tuning was used, respectively.

We start by exploring the approach using the linguistic *CFGR* features in Section 14.4.1. Then, we investigate optimizing the n-gram models in Section 14.4.2, before we combine both in Section 14.4.3. For *CFGR* and n-grams, we conduct both, single- and cross-corpus experiments utilizing T11 and NT11 data (see Section 14.1) in order to provide a better assessment of the ensemble approach with

[2]This means that we also include the two n-gram features from Chapter 6. On the one hand they consist of combined n-grams of different lengths, thus technically constituting models, different from the 40 individual ones constituting our basis. On the other hand, we opted for using all of the 16 features for consistency reasons. In the end, it is up to the model selection algorithm to decide on the use of the individual features.

[3]See the *Tools* sections in the corresponding chapters for details.

surface and some more linguistically-motivated features. Finally, we focus on the T11 data and investigate, whether it is possible to outperform the best published results on this standard data set, by adding some more of the explored linguistic features to the ensembles.

14.4.1 CFGR

Here we utilize the variationist features based on CFG rules (*CFGR*), which we introduced and explored in detail in Chapter 9.2, and systematically explore the ensemble optimization and tuning procedure discussed in Chapter 13.

Ensemble results for the CFGR variables In Chapter 9.2, we explored non-lexicalized ($CFGR_{ph}$) and lexicalized ($CFGR_{lex}$) features based on CFG rules in detail, including investigating the performance of classifiers based on individual syntactic categories as variables (i.e., separate classifier for *VP*, *NP* and *S*, etc. as well as *NN*, *VB* and *JJ*, etc. as variables), using the best performing variationist encoding var_w. Here, we follow on this, and explore our ensemble approach using these individual variable-based models. We include all of them, i.e., the union of $CFGR_{ph}$ and $CFGR_{lex}$ constituting the $CFGR_{ph \cup lex}$ feature type.

The ensemble results for the separate variable-based models are presented in Table 14.1. The column *baseline* lists the corresponding results from Table 9.1 (see Section 9.2.4), which were obtained by putting all the features in a single vector (i.e., no ensemble is used to combine the different feature types). The number in parentheses specifies the number of models combined in the ensemble: in the *features* column, it shows the overall number of separate variable-based models, and in the +*opt* columns, it is the number of models selected by the optimization algorithm.

The results show that generating an ensemble using all of the individual variable-based models without optimization and tuning *(-opt, -all)* leads to a notable drop in accuracy compared to the baseline. The fact that the drop in the cross-corpus setting is more than 20% is particularly striking. We assume that this is due to the poor performance of most of the individual models, yielding probabilities of little use overall. The few relatively well-performing models we discussed in sec-

features	data	baseline	ensemble			
			-opt		+opt	
			-all	+all	-all	+all
$CFGR_{ph \cup lex}$ (71)	sc	78.8%	66.0%	79.2%	71.3% (14)	**79.6% (8)**
	cc	38.8%	18.1%	34.2%	32.6% (10)	**39.0% (1)**

Table 14.1: Results for the $CFGR_{ph \cup lex}$ ensembles with different optimization settings

tion 9.2.4 apparently are flooded by the noise introduced by the others. Thus, for a set of rather low-performing models without any optimization, it seems preferable to provide the classifier with access to the individual features instead of to the noisy probability estimates. The optimization *(+opt, -all)* leads to a clear performance improvement over the non-optimized settings. In the single-corpus setting only 14 of the 71 models were kept and in cross-corpus only 10.

Table 14.2 shows the selected models in the order in which they are selected by the ensemble optimization algorithm. For *(+opt, -all)*, the table essentially consists of the best performing variables (see Section 9.2.4), suggesting that the algorithm makes meaningful choices.

data	$CFGR_{ph \cup lex}$: selected models	
	+opt, -all	+opt, +all
sc	[NN]+[JJ]+[RB]+[NNS]+[VB]+[NP]+[S]+[VP] +[IN]+[VBP]+[VBG]+[VBN]+[NNP]+[,] (14)	[*all*]+[NN]+[JJ]+[RB]+[PRP]+[VBN]+[NNP]+[WDT] (8)
cc	[NN]+[JJ]+[NNS]+[NP]+[RB]+[VB]+[VP]+[NNP] +[S]+[IN] (10)	[*all*] (1)

Table 14.2: The $CFGR_{ph \cup lex}$ model sets selected by optimization

The flipside of the coin is that low-performing models generally were not found to have a positive effect and thus were not included. Yet, optimization by itself is not successful overall, given that the *(+opt, -all)* accuracy remains below the single feature set baseline.

Applying tuning without optimization *(-opt, +all)* outperforms the optimization result. Thus, including the overall model [*all*] in the ensemble improves the meta-classifier. In the single-corpus setting, the accuracy is slightly higher than the baseline, in cross-corpus it remains below the baseline.

Turning on both, optimization and tuning *(+opt, +all)*, yields the overall best

results of Table 14.1 – 79.6% for single-corpus and 39% for the cross-corpus setting. The corresponding entry in Table 14.2 shows that tuning significantly reduces the number of selected models. This is not unexpected given that the overall model [*all*] essentially includes all the information. In the cross-corpus setting, [*all*] indeed is the only model selected. Interestingly, in the single-corpus setting, the optimization algorithm identifies some additional models to improve the accuracy, mainly ones that also perform well individually. While this amounts to adding information that in principle is already available to the [*all*] model, the improvement may stem from the abstract nature of the probability estimates used as features of the meta-classifier. When both, optimization and tuning, are applied, the tuning apparently stabilizes the ensemble leading to higher performance, and the optimization algorithm further improves the result by reducing the noise.

14.4.2 Recurring N-grams

In the previous section, we showed that the linguistic *CFGR* features can well benefit from our ensemble optimization and tuning approach. As pointed out in Section 14.2, combining the more linguistically-motivated features with n-grams as basis, seems attractive in terms of NLI performance. Before we turn to exploring such combinations, we first investigate applying our ensemble approach to n-grams as well. It might help reducing the complexity of the overall system by identifying a potentially well-performing subset of the 40 n-gram models.

Table 14.3 provides the results for the n-gram ensembles built on the basis of the recurring word-, lemma-, POS-, OCPOS-based n-grams with $1 \leq n \leq 10$ (see Section 14.2) in the same format as Table 14.1 for $CFGR_{ph \cup lex}$. Different from the $CFGR_{ph \cup lex}$ case, the results for the n-gram ensemble model without optimization or tuning *(-opt, -all)* already are 4–5% higher than the single vector baseline.

The best results, 83% for single-corpus and 36.5% for the cross-corpus setting, are obtained by applying the optimization. The n-gram ensembles seem to benefit more from optimization than from tuning in general. The feature counts for the n-grams (single-corpus: 4,822,874; cross-corpus: 3,687,375) are far higher than for $CFGR_{ph \cup lex}$ (single-corpus: 98,115; cross-corpus: 94,176), so there may be

features	data	baseline	ensemble			
			-opt		+opt	
			-all	+all	-all	+all
N-GRAMS (40)	sc	77.1%	82.3%	82.6%	**83.0% (13)**	82.3% (8)
	cc	31.0%	34.9%	34.6%	**36.5% (6)**	35.5% (6)

Table 14.3: Results for the n-gram ensembles with different optimization settings

more noise in the [*all*] model, making it less useful for the tuning step.

Table 14.4 lists the models selected by the optimization algorithm in order in which they are selected. The n-gram types and the n of the model is indicated, e.g., *[OP-3]* means *OCPOS-based trigrams*[4]

data	N-GRAMS: selected models	
	+opt, -all	+opt, +all
sc	[W-2]+[L-2]+[W-1]+[L-1]+[L-3]+[W-3]+[OP-3] +[OP-1]+[OP-5]+[P-3]+[P-5]+[P-2]+[OP-8] (13)	[*all*]+[W-2]+[L-2]+[W-1]+[L-1]+[L-3]+[OP-4]+[L-4] (8)
cc	[W-2]+[W-1]+[L-1]+[L-3]+[W-3]+[OP-2] (6)	[W-2]+[W-1]+[*all*]+[L-1]+[L-3]+[P-4] (6)

Table 14.4: The n-gram model sets selected by optimization

For the more surface-based n-gram (word- and lemma-based), the optimizer selected only up to $n = 3$, whereas for the more abstract ones (POS- and OCPOS-based), models up to $n = 8$ were included. Thus, different from the variables-based $CFGR_{ph \cup lex}$ ensemble, we here find that individually relatively poorly performing models such as those considering longer n n-grams (cf. Chapter 5), are kept when optimizing the ensemble. Complementing our findings in Chapter 5, the outcomes suggest that when abstracting from the surface, one can get some useful information out of longer n-grams that apparently is not contained in the short surface-based ones.

14.4.3 Combining Recurring N-grams and CFGR

After having explored our ensemble approach using the *CFGR* and n-gram features in detail, we can turn to combining both.

The results are presented in Table 14.5. We explore four different ways to combine the two model sets, and the table shows the best results for each of the

[4]See the shorthand notation for n-grams introduced in Section 14.2.

features	data	ensemble			
		-opt		+opt	
		-all	+all	-all	+all
(a) $CFGR_{ph\cup lex}$ (71) + N-GRAMS (40)	sc	82.1%	82.9%	82.9% (20)	**83.6% (6)**
	cc	34.1%	36.0%	36.7% (8)	**38.5% (3)**
(b) $CFGR_{ph\cup lex}$ (71) + N-GRAMS [+opt, -all] (ME)	sc	83.1%	83.7%	82.6% (4)	**84.2% (5)**
	cc	37.4%	39.6%	38.0% (3)	**40.3% (3)**
(c) $CFGR_{ph\cup lex}$ (+opt, +all) (ME) + N-GRAMS (40)	sc	83.7%	**84.8%**	84.7% (13)	83.8% (13)
	cc	36.8%	38.9%	42.0% (5)	**43.0%** (4)
(d) $CFGR_{ph\cup lex}$ (+opt, +all) (ME) + N-GRAMS (+opt, -all) (ME)	sc	**83.5%**	**83.5%**	**83.5% (2)**	83.4% (2)
	cc	41.3%	**42.0%**	41.3% (2)	40.6% (2)

Table 14.5: Optimization results combining n-grams and $CFGR_{ph\cup lex}$

setups in bold, once for the single-corpus and once for the cross-corpus setting. In the following we explain the different setups in detail.

- *Setup (a)*: For the results of setup (a), we use the ensemble consisting of all individual models separately.
- *Setup (b)*: In (b), the $CFGR_{ph\cup lex}$ models are included as in (a), but we replace the n-gram models by a *single meta-ensemble model (ME)* generated using the best n-grams setting *(+opt, -all)*, which consists of 13 models for single-corpus and six models for the cross-corpus setting (see Table 14.3). ME thus is a *meta-meta-classifier*, generated by applying the ensemble model generation routine to an ensemble.
- *Setup (c)*: In (c), we invert the (b) setting: The $CFGR_{ph\cup lex}$ features are replaced by a meta-ensemble generated using the best performing $CFGR_{ph\cup lex}$ setting *(+opt, +all)*, which consists of eight models for the single-corpus, and one model for the cross-corpus setting (see Table 14.1).
- *Setup (d)*: Finally, in (d) we combine the meta-ensemble for $CFGR_{ph\cup lex}$ with the meta-ensemble for the n-grams obtaining an ensemble consisting of two models.

The best results of 84.8% in the single-corpus setting and 43% cross-corpus, underlined in the table, are obtained in setup (c). These are the overall best results

across all experiments described in this thesis so far. The best result in the single-corpus setting involves tuning only, whereas in the cross-corpus setting it involves tuning and optimization selecting the models [*all*]+[$CFGR_{ph \cup lex}$ (+opt, +all)]+[W-2]+[W-1]. This again shows that both, surface and linguistically-motivated features are important for developing robust, high performing NLI systems.

The single-corpus accuracy of 84.8% outperforms the best result of the First NLI Shared Task by 1.2%, using the same standard T11 data setup (see Section 3.2). In the cross-corpus setting, the 43% accuracy also outperforms the previous best result on the NT11 data by 4.5% (see Chapter 6).

In sum, the overall best results in the single-corpus and cross-corpus settings are obtained starting with the whole n-gram model set plus an optimized $CFGR_{ph \cup lex}$ meta-ensemble. This confirms the usefulness of the optimized ensemble setup and underlines that combining a range of linguistic properties, from n-grams at different levels of abstraction to local syntactic trees characteristics, is a particularly fruitful approach for NLI as a good example of an experimental task putting linguistic modelling to the test with real-life data.

14.4.4 Maximizing Performance Using Linguistic Features

In Section 14.4.3, we showed that combining n-grams with the more linguistically-motivated *CFGR* features, yields a performance in line with the state-of-the-art. Here, we focus on the standard single-corpus T11 data setup (see Section 14.1), and investigate, whether it is possible to further improve the accuracy by combining different linguistic features into ensembles.

The results are summarized in Table 14.6. The table shows various systems, ranked by accuracy, along with the corresponding setup ids, the type of the model (ensemble or simple), the number of models used in ensembles, and a reference (for models taken from our previous explorations). Besides our own systems, the table also lists Ionescu et al. (2014) as (c), which is the best performing system reported so far by other researchers. In the following we describe the various setups. We start with the separate models selected for our experiments, and then turn their combinations using our ensemble approach.

rank	id	system	type	# models	reference	accuracy
1	(a)	**(b) + (i)**	E (-opt, +all)	58 + 1	-	**85.5%**
2	(b)	**(d) + (j)**	E (-opt, +all)	57 + 1	-	**85.4%**
3	(c)	*Ionescu et al. (2014)*	S	-		85.3%
4	(d)	(e) + (f)	E (-opt, +all)	41 + 16	-	85.2%
5	(e)	(g) + (h)	E (-opt, +all)	40 + 1	Sec. 14.4.3	84.8%
6	(f)	FNLIST	E (-opt, -all)	16	Ch. 6	82.5%
7	(g)	N-GRAMS	E (-opt, -all)	40	Sec. 14.4.3	82.3%
8	(h)	CFGR	ME	1	Sec. 14.4.3	79.6%
9	(i)	SOX-POS-NoP, best (dev)	S	-	Sec. 9.3	49.1%
10	(j)	SOX-NoP, best (dev)	S	-	Sec. 9.3	48.0%

Table 14.6: Various systems (E: ensemble, ME: meta ensemble, S: simple)

- *Setup (g)*: We use the separate 40 n-gram-based models as basis (see Section 14.2). In order to have a most flexible basis, we do not optimize or tune this set[5] – Depending on which features get added to the basis, different of the individual n-gram models might be useful. We let the optimization procedure determine the best basis.

- *Setup (h)*: The *CFGR*, or more precisely $CFGR_{ph \cap lex}$, feature type consists of 71 models, each representing a particular mother node or POS category as variable. This number of models is quite high for representing a single feature type, and would significantly increase the complexity of the overall ensemble. Thus, we employ the optimized and tuned version *(+opt, +all)*, which reduces the number of models from 71 to 8, and then we further reduce the resulting 8 models into a single meta-ensemble (ME) model as described in Section 14.4.3. In sum, the 71 original models are thus represented by a single meta-ensemble model, which we utilize for the different ensemble setups explored in this section. The findings in Section 14.4.3 show that this seems to be a reasonable way to deal with the given feature type – Combining it with with n-grams as described in (g), yielded our so far best performing system[6], which is represented here by the setup (e).

[5]Ensemble parameters: *(-opt, -all)*

[6]See Section 14.4.3, in particular the Table 14.5, setup (c).

- *Setup (f)*: This setup represents the broad feature set, consisting of the 16 feature types utilized in Chapter 6, i.e., explored in the context of the First NLI Shared Task (FNLIST). Since it includes very different feature types, we do not optimize or tune this set[7], but employ all of the individual models.

- *Setup (i)*: This setup represents our most specific *SOX* feature type, namely *SOX-POS-NoP*, discussed in Section 9.3. We use the best performing version, namely [*s/c, +bvm*] at cut-off 0.0, with the cut-off parameter tuned on the *dev* set.[8]

- *Setup (j)*: This setup represents our most abstract *SOX* feature type, namely *SOX-NoP*, explored in Section 9.3. We use the best performing version, namely [*s/c, +bvm*] at cut-off 0.7, with the cut-off parameter tuned on the *dev* set.[9]

After having described the separate models, we turn to combining them following our ensemble approach.

- *Setup (e)*: This setup reflects an ensemble combining the n-grams (g) and the *CFGR* meta-ensemble (h). As mentioned above, it constitutes our best performing system thus far in this thesis (see Section 14.4.3). The best parameter setting for this ensemble is *(-opt, +all)*.

- *Setup (d)*: This setup constitutes an ensemble combining (e) and (f), i.e., here we extend our so far best performing model (e) by the 16 features types used in Chapter 6 (f). It shows an improvement by 0.4% compared to (e) alone, and thus constitutes our new best performing system in this thesis. The best parameter setting for this ensemble is *(-opt, +all)*.

- *Setup (b)*: Here, we extend (d) by (j), i.e., our new best performing system (d) by the *SOX-NoP* feature type explored in Section 9.3 (j). It shows a

[7]Ensemble parameters: *(-opt, -all)*

[8]We tune the cut-off parameter on the T11 *dev* set while training on the T11 *train* set here, avoiding any tuning on the *test* set (which can be misleading). On the *dev* set, the best parameter turned out to be 0.0, whereas on the test set it was 1.0 (see Section 9.3).

[9]Again, we tune the cut-off parameter on the T11 *dev* set while training on the T11 *train* set here. Interestingly, the best cut-off turned out to be exactly the same as on the T11 *test* set (see Section 9.3).

further improvement by 0.2%, outperforming our so far best system in this thesis (d), as well as the best system reported thus far by other researchers on the given data setup (c). The best parameter setting for this ensemble is *(-opt, +all)*. This system is part of our work in Bykh & Meurers (2016) – To the best of our knowledge, it shows the best accuracy published so far on the standard T11 data setup.

We also wanted to investigate the contribution of the feature grouping technique proposed in Section 8.3 in the context of a comprehensive ensemble model as used here. Thus, alternatively we added the basic *SOX-NoP* setting [*s/c, +bvm*] *without* clustering, to the ensemble (d), instead of adding the best performing setting incorporating grouped variables (IGVs) as represented by (j). Both options showed an increase in accuracy by the same value of 0.2%. In other words, there was no quantitative difference between adding the best performing [*s/c, +bvm*] setting (cut-off 0.7), which incorporates complex features (IGVs), and the lower performing version containing simple features (FGVs) only. Yet, there is a difference in terms of the feature counts for the two models, namely, 16,841 vs. 18,174 respectively. Thus, the version incorporating IGVs is more efficient, providing the same quantitative advantage with a more compact model. This supports the assumption that the generalizations made by the technique are reasonable. The findings suggest that the approach can further advance and optimize the state-of-the-art NLI systems.

- *Setup (a)*: Here, we extend (b) by (i), i.e., our new best performing system (b) by the *SOX-POS-NoP* feature type explored in Section 9.3 (i). It shows an accuracy of **85.5%**, i.e., a further improvement by 0.1%, thus constituting the **overall best system** in this thesis and published so far based on the standard T11 data setup. Again, the best parameter setting is *(-opt, +all)*.

In sum, combining a range of features explored in the course of this thesis utilizing the proposed ensemble approach, we obtained an accuracy of 85.5%, which is the best result published so far using the standard T11 data setup.

14.5 Conclusions

In this chapter, we aimed at further advancing the performance employing a wide range of feature types, explored in the course of this thesis. For this, we utilized meta-classifiers, namely probability-based ensembles, and proposed a method for optimizing and tuning them. It turned out that the suggested optimization and tuning procedures are well-capable of advancing the results. Combining surface-based and different linguistically-motivated features explored in the course of this thesis, showed an accuracy of 85.5%, outperforming the best systems reported so far using the standard T11 data setup.

Connecting the findings and the research questions Regarding the particular five research questions in the focus of this thesis (see Section 1.3), the discussion in this chapter contributes to four of them as follows:

1. [LINGUISTIC-FEATURES]: We combined a wide range of surface-based and linguistically motivated features, explored in the different chapters of this thesis. The results confirm our findings in Chapter 6. On the one hand, the outcomes further exemplify that surface-based features such as different n-grams constitute an important classifier basis, providing a reasonable classification accuracy for NLI. On the other hand, the results show that adding linguistic features on top of the surface-based models, can indeed further improve the already high accuracies. Thus, linguistic feature engineering is not only interesting from the theoretical perspective, but also seems to be an important part in developing high performing state-of-the-art NLI systems.

2. [VARIATIONIST-PERSPECTIVE]: Incorporating variationist features explored in Chapter 9 into the overall NLI system consisting of a range of different feature types, showed further performance increase, outperforming best results reported so far. Thus, variationist features are not only useful in qualitative explorations advancing the SLA insight as shown in Chapter 10, but they indeed also provide a quantitative edge, and can further advance high-performing state-of-the-art NLI systems.

3. [MODEL-OPTIMIZATION]: This chapter was mainly dedicated to combining a broad range of features into single NLI systems, with the aim to further improve the classification performance. This raises the question of how to do this combination, and how such huge models could be optimized and tuned. For this, we employed a meta-classifier approach. In particular, we used probability-based ensembles, which were shown to outperform models, simply combining all features into a single vector. Further, we proposed a method for optimizing such ensembles by utilizing a linear model selection algorithm. The procedure seems well-capable of selecting an appropriate subset of the models, showing high accuracies, often outperforming ensembles utilizing the whole model set. Moreover, we suggested a tuning procedure which is based on including all of the available features in an abstract, condensed form into the overall ensemble. This information, which is apparently redundant, turned out to be still very useful for the ensemble classifier – Indeed the largest, most diverse models benefited more from this tuning than from the optimization procedure (see Table 14.6). We assume that the reason why optimization was less useful than tuning for large, diverse models is that dropping models from a very diverse sets, can lead to a critical information loss. Whereas the proposed tuning does not drop any information – On the contrary, it provides some additional information to the system. For several settings, the best performance was achieved by first employing tuning and then optimization. Apparently, the tuning stabilizes the ensemble (i.e., makes it more robust) leading to better performance, and the optimization algorithm can then optimize the model by reducing the noise or redundancy. Thus, the interplay of both components can yield optimized models, which on the one hand, are less complex in terms of the actual feature counts, and on the other hand, can show an improved performance compared to the simple ensembles. In sum, our ensemble approach seems to be a useful tool for optimizing large NLI systems. Moreover, we showed that the label-informed feature grouping technique presented in Section 8.3, can also contribute to optimizing models by reducing the feature space while retaining the accuracy level, thus making the models more compact and efficient.

4. [CROSS-CORPUS]: As usual, our cross-corpus results were lower than the single-corpus outcomes. However, our ensemble approach showed comparable benefits for both, single- and cross-corpus settings. It seems to generalize across data sets, and thus to be of general use. Different from our findings in Chapter 6, extending a system based on n-grams by some more linguistically-motivated features also improved the cross-corpus accuracies in our experiments. Thus, the success in this regard seems to depend on the actual feature types and classifications options.

Part V

Conclusions

Chapter 15

Summary and Outlook

In this thesis, we explored the task of NLI from both, quantitative as well as qualitative perspectives. First, we showed a way to design state-of-the-art NLI systems yielding accuracies over 85% for a standard data setup with a chance baseline of 9.1%, outperforming previously published results. To achieve this, we utilized different surface-based features, engineered a diverse linguistically-motivated feature set, and combined the different models using meta-classifier techniques. Second, we showed how a qualitative analysis based on NLI outcomes can foster the discovery of new findings in the area of SLA, and thus advance linguistic insight.

This chapter presents the conclusions and contributions of this thesis in more detail. In particular, Section 15.1 provides part- and chapter-wise conclusions. Section 15.2 summarizes our findings with respect to the particular research questions explored in this thesis (see Section 1.3). Section 15.3 discusses the limitations of the presented work and sketches the outlook.

15.1 Summary

Part I (Chapters 1–3) clarifies the context of this thesis, presents the research questions and discusses the related work on NLI. Parts II–IV (Chapters 4–14) describe our study, and present our results and findings which we summarize part- and chapter-wise in the following sections.

15.1.1 Broad Linguistic Feature Exploration

Part II describes our results utilizing a broad range of surface-based and linguistically-motivated feature types. Chapter 4 presents a brief introduction and outline for this part, while the Chapters 5–6 present the actual findings, which are summarized in more detail below.

Recurring n-grams as features for NLI In Chapter 5, we explored the use of *recurring n-grams* of different types as features for NLI. In particular, we extracted word-, Part-of-Speech- (POS) and Open-Class-POS-based n-grams of any length, re-occurring in the training data, and explored their performance in a Machine Learning setup. Our system showed a high accuracy of 89.7% in a single-corpus setting utilizing seven L1s from the ICLE corpus. N-grams up to the length of $n = 5$ turned out to be useful, thus it might be advantageous to consider higher n than is usually done in the related research. Combining n-grams of different lengths showed a better performance than using single n n-grams.

Furthermore, our system also showed a reasonably high accuracy of about 87.6% in a cross-corpus setting employing texts for three L1s drawn from ICLE, HKUST, USE and NOCE corpora. The single-corpus results with the same three L1s based on ICLE were at 96.5%. Thus, given a drop of only 9% between single- and cross-corpus outcomes, our results suggest that patterns learned on ICLE can generalize to other data sets. This finding is contrary to Brooke & Hirst (2011), reporting much worse cross-corpus results using ICLE and Lang-8 data. We assume that the better performance of our system can be attributed to a higher similarity of the data used in our experiments regarding the general structure and the genre.

In general, the more abstract n-grams incorporating POS performed worse than the pure surface-based word n-grams in our core study. However, in a separate exploration, we showed that linguistic abstraction via POS can provide quantitative advantages though, especially if applied to n-grams of higher n, combining word and POS information.

Surface-based and linguistically-motivated features and feature combinations In Chapter 6, we explored using a broad range of feature types in the context of

the *First NLI Shared Task* (Tetreault et al., 2013). We considered surface features, namely, recurring word-based n-grams investigated in Chapter 5, as our basis. We then explored the performance of different linguistically-motivated feature types. Some of them are novel for the task of NLI, such as recurring function based dependencies, dependency realization features, or different suffix-based features. Others have been previously employed for NLI. In sum, using a probability-based ensemble classifier combining features based on POS, dependency and constituency trees as well as lemma realization, complexity and suffix information, we were able to outperform the already high accuracy achieved by the word-based n-grams alone. Thus, the exploration of different linguistically-informed features is not just of conceptual interest, but also provides an advantage in terms of classification accuracy.

Furthermore, our findings suggest that it seems possible to optimize ensemble classifiers by model selection instead of combining all of the available models. It could reduce the noise or eliminate potentially redundant information from the feature space, and thus, make the systems less complex and more efficient.

Our findings suggest that the following recipe seems promising in developing high-performing NLI systems:

1. Start with surface features, such as word-based n-grams as basis.

2. Add some more elaborated linguistically-motivated features, capable of capturing potential L1-transfer effects at different linguistic levels (e.g., morphology, syntax, etc.)

3. Optimize the model, e.g., using model selection techniques.

15.1.2 A Variationist Approach to NLI

Part III (Chapters 7–11) presents our findings on applying a particular linguistic theory, namely, a *variationist sociolinguistics perspective* to the task of NLI.

General consideration on applying a variationist perspective to NLI In Chapter 7, we discussed general issues in applying the variationist perspective to the task of NLI. At the heart of the approach is the notion of a *linguistic variable*

representing some linguistic concept or structure, and which can be realized by a set of options called *variants*. We assume the following potential benefits in connection with this approach:

1. Obtaining *quantitative* advantages based on potentially highly indicative preferences in the non-native language productions by writers with a different L1.

2. Fostering the *qualitative* analysis by unveiling the particular indicative choices, which then can be considered and interpreted within the linguistic theory.

The guiding questions for this part can be summarized as follows:

1. What is a suitable definition of a *linguistic variable*?

2. What different *types of linguistic variables* should be distinguished in the given context?

3. How can we *abstract over individual linguistic variables* to obtain insights into more general underlying linguistic structures reflected in NLI?

4. What *linguistic variables to explore* as features in the context of this thesis?

Giving answers to these questions constitutes subsequent steps in the features engineering process, described and implemented in Chapters 8–10.

Variationist feature engineering In Chapter 8, we revised the notion of the linguistic variable, and based on previous research, we proposed a broader definition of this notion, which we consider more suitable in the context of this thesis, and for related tasks in general. In particular, the proposed definition relaxes the traditional semantic constraint of meaning equivalence, which seems rarely sustainable in the strict sense beyond the phonetics-phonology level anyway. Moreover, exploring preferences for certain meanings conveyed by particular forms poses a highly interesting and promising direction, worthwhile pursuing in the variationist sociolinguistics research in general, and in the context on NLI in particular.

Furthermore, we suggested a taxonomy of linguistic variables, useful for our study. We distinguish between relative vs. absolute variables, lexical vs. grammatical variables, as well as between variables at different levels of granularity.

Moreover, we proposed a machine learning technique based on hierarchical clustering for implementing variables of different granularity, which we refer to as label-informed feature grouping. The technique can be used to abstract over some individual variables by grouping those variables together which behave alike with respect to the classification label (L1), thus providing a more general perspective.

We evaluated and exemplified our approach from the quantitative and qualitative perspectives in the subsequent chapters.

Quantitative explorations of the variationist approach In Chapter 9, we explored several linguistic features under a variationist perspective, focusing on the quantitative aspect. We exemplified the taxonomy presented in Section 8.2, and showed how linguistic variables of different types can be generated based on the principles of variationist sociolinguistics. In particular, we investigate two specific sorts of such variables, which on the one hand are related in that both are situated at the syntax level, but on the other hand largely differ with respect to the proposed taxonomy, namely, variables related to syntactic category realization and variables reflecting verb subcategorization.

Our findings suggest that using feature encodings based on the variationist perspective can indeed provide a quantitative edge, compared to simple frequency or binary representations. Furthermore, we applied the proposed feature grouping technique to suitable linguistic variables and explored the quantitative potential of the method. It turned out that the grouping technique can provide quantitative advantages, especially if applied to more abstract linguistic features and supported by surface models. It seems capable of generating reasonable groups of variables and to some extent overcome data sparsity issues.

In referring to the provocative question in the title of this thesis regarding the choice in the non-native productions, we can state the following: Given the high accuracies up to 79% (given a chance baseline of 9.1%), the learner choices and preferences for certain variants realizing particular linguistic variables seem to be definitely influenced by the L1. However, there are clearly many factors

such as proficiency, social status, age or gender, etc., which certainly may also influence the choices to an extent which is not clear at this point. Unfortunately, controlling for all of them seems hardly feasible with the given data. Nevertheless, our results provide a piece to the overall puzzle constituting the answer to this thrilling question.

Qualitative explorations of the variationist approach The high performance obtained in the quantitative part of our explorations, is an important indicator for features, capturing potentially interesting differences in the language use by learners with different L1s. However, identifying high performing settings can only be the first step in the overall procedure of advancing the scientific insight. The second step should be the qualitative analysis of the outcomes. This means that we have to employ suitable methods, and to identify and motivate particular interesting L1-transfer candidates.

Thus, in Chapter 10, we focused on the qualitative aspect of our variationist approach to NLI. For our explorations, we employed subject realization and nominal modification features, conceptually well-motivated by the related work and well-suitable in the given context. We proposed and discussed a comprehensible method capable of discovering interesting L1-transfer candidates, utilizing logistic regression weights assigned to linguistic variables at different levels of granularity (see Section 8.2.3). On the one hand, we exemplified a procedure for discovering interesting L1-transfer candidates in a data-driven way. In particular, we showed how to obtain first promising findings using the most coarse-grained variable type according to our taxonomy, and how using different settings employing more fine-grained variables, these findings can be gradually refined leading to valuable new insights. On the other hand, we showed how our approach can foster the theory-driven analysis by facilitating the validation of existing and new hypotheses about L1-transfer. In particular, we tested the hypothesis by Wang (2009) regarding the preferences in the subject realization by L1 Chinese learners of English as L2. Our method employing the proposed feature grouping technique was capable of discovering specific features, supporting and not supporting this hypothesis, which allows for further fine-grained analyses. The method seems to provide a fruitful integration of machine learning techniques and a variation-

ist perspective. Our findings suggest that it is well capable of contributing new insight on L1-transfer, and thus advancing the SLA research.

Our qualitative findings contribute another piece of evidence to the clarification of the question in the title of this thesis, suggesting if there is a real choice in non-native productions, or if it is maybe predetermined by the L1 of the learners. In particular, we discovered that the variants *there is* vs. *there exists*, are both used by L1 German learners in L2 English apparently as equivalent to the frequent German *es gibt*, whereas *there is* is a regular realization and *there exists* is rather unusual. Thus, one could argue that, in a way, learners even have a *"greater choice"* in their non-native voice than the natives in their productions. More precisely, the learners show regular and various irregular realizations (some of them apparently due to possible L1-transfer effects), and thus can realize an extended set of options, compared to the natives generally realizing regular forms. Identifying distinctive choices and preferences in the regular realizations as well as indicative irregular cases remains a highly interesting and thrilling research direction. We hope that our method will pose a useful tool for discovering new valuable findings on that point in the future work.

Advantages and limitations of the variationist approach In Chapter 11, we discuss the advantages and limitation of our variationist approach to NLI.

Advantages

- *Conceptual advantages*: The approach allows for a specific view on language units at different linguistic levels, i.e., it allows for considering and exploring a set of conceptually, structurally or contextually related units in direct connection to each other. Furthermore, it makes possible to infer new, more general linguistic variables based on the original set, in a linguistically-informed way. For this, we proposed a flexible machine learning method based on hierarchical clustering, which we refer to as label-informed feature grouping.

- *Quantitative advantages*: Variationist features are capable of providing practical advantages in terms of classification accuracy. Given variationist fea-

ture encodings, the classifier can figure out most indicative variant distributions for different L1s, which can lead to an improved performance.

- *Qualitative advantages*: Variationist features allow to concentrate on the proportions of related variants in the data, which can help discovering new interesting usage patterns for learners with different L1s in a data-driven way, enhancing the theoretical insight. Moreover, such a view on the features allows for formulating and testing various theory-driven hypotheses about L1-transfer, assuming different variant preferences for particular linguistic variables across learners with different L1s. One can also combine the data- and theory-driven procedures: Identify interesting variant choices in a data-driven way first, and then use this information to formulate and validate new theory-driven hypotheses about L1-transfer.

Limitations

- *Conceptual limitations*: So far the method does not provide any means for ensuring semantic constraints, such as the traditional meaning equivalence requirement, usually imposed in variationist sociolinguistics studies. While we argue that always keeping such constraints is not necessarily beneficial in any context, in fact sometimes posing an unnecessary restriction, we believe that extending the approach by some semantic components could be an interesting direction worth pursuing in terms of the future work though.

- *Quantitative limitations*: There are certain limitations in connection with the proposed label-informed feature grouping technique aimed at automatically generating linguistic variables at more abstract levels. While feature grouping as proposed here, is capable of reducing data sparsity by considering groups of related variables instead of individual (possibly rare) ones, it apparently also hides some specific traits of the individual variables which can be indicative and thus useful in terms of classification accuracy. Indeed, we showed that adding some surface-based models to the system, which are capable of restoring the cues glossed over by grouping, leads to an improved performance. Further, the more specific the features become, the

less seem the advantages of the grouping, even if its supported by corresponding surface models: It is apparently more suitable for the classifier to have the access to the individual specific features, and to weight them separately, which increases the chance of spotting highly indicative individual combinations of variables and variants not accessible after the grouping. In addition, the performance gains of around 2%, we were able to obtain, are not as high as expected. However, we showed that this could be attributed to the rather limited size of the standard data set used in our explorations. In sum, grouping proves most advantageous if it gets applied to some more general and abstract linguistic features, and at the same time gets supported by appropriate surface models.

- *Limitations of the qualitative analysis*: The more abstract linguistic variables induced by the proposed automatic feature grouping method, pose some difficulties in terms of qualitative interpretation. However, this does not mean that a meaningful interpretation is not possible though. The members of some groups show certain common properties. It might require more in depth analyses in order to identify the specific properties making those groups indicative for particular L1s. Another point is that the indicative power of the features in the employed settings, always means that these are indicative in relation to the other L1s in the set. Thus, depending on the data, the outcomes might differ. However, since we employed a data set that includes a reasonable number of very different L1s, we believe that our findings are conclusive though. Further, in the qualitative analysis we consciously avoided any strong conclusions about L1-transfer, speaking about L1-transfer *candidates* instead. There are three main reasons for this:

    1. The employed data, namely TOEFL11 (T11), was compiled specifically for the task of NLI, and it seems generally to be well suitable for the given task. However, we believe that it is not sufficiently large for any strong claims about L1-transfer though. Some of the features show rather low frequencies, especially the more specific ones. We hope that, e.g., the more comprehensive TOEFL-Big corpus (Tetreault et al., 2012) will be made publicly available in the future, which cer-

tainly would be helpful in this context.

2. The learner data apparently poses a challenge for standard NLP tools, leading to tagging and parsing errors. Using tools which are better tuned to this specific language variety, would certainly help reducing these issues.

3. Third, we believe that any findings should be verified using a range of appropriate corpora, applying different suitable methods, before making any strong conclusions.

In practice, it is rather difficult to meet all of these requirements. Nevertheless, we believe, this is necessary in order to ensure high quality of the findings, and thus should be of high priority in the future work.

15.1.3 Advancing Performance

Part IV is dedicated to further advancing the performance of NLI systems. Especially, it targets the question of how to combine a range of different features, and how such complex models can be optimized and tuned. Chapter 12 presents a brief introduction and outline for this part. The Chapters 13–14 describes our approach and its evaluation respectively.

Ensemble optimization and tuning approach In Chapter 13, based on previous research and our findings in the course of this thesis, we propose using a meta-classifier approach for combining different features types into a single system. In particular, we propose using a probability-based ensemble as suggested by Tetreault et al. (2012) and explored in Chapter 6. Furthermore we present ensemble optimization and tuning techniques capable of further advancing the models. The optimization is implemented by a linear model selection algorithm. Tuning is implemented by adding an abstract model, which incorporates the whole feature set in an abstract form into the ensemble. We believe that it can be helpful in dealing with increasing numbers of features employed for NLI. On the one hand, tuning can lead to an improved performance by stabilizing the ensemble. On the other hand, optimization could help reducing the complexity of the systems and

the noise in the feature space by suggesting appropriate subsets of the available models. We were particularly interested in the interplay of both. The method was implemented and evaluated in Chapter 14.

Advancing performance using ensembles In Chapter 14, we explore combining a broad range of features into single NLI systems, with the goal of further improving the classification accuracy. For that, we implement and evaluate our ensemble approach presented in Chapter 13, utilizing the proposed optimization and tuning method.

- *Ensemble optimization*: The proposed ensemble optimization procedure employing a linear model selection algorithm, turned out to be capable of selecting reasonable subsets of the models, showing high accuracies. Some of the generated sub-systems outperformed the overall system.

- *Ensemble tuning*: The suggested ensemble tuning procedure, which is based on including all of the available features in an abstract, condensed form (i.e., as just a single individual model, represented by another probability distribution) into the ensemble, showed reasonable quantitative advantages.

- *Combining ensemble optimization and tuning*: Indeed the largest, most diverse models benefited from our tuning method more than from the optimization. Presumably this is due to the fact that dropping models from such diverse systems can lead to a critical information loss in the end. Whereas the proposed tuning does not drop any information – On the contrary, it provides some additional information to the system. However, for several single- and cross-corpus settings, the best performance was achieved by employing both techniques, i.e., tuning and optimization in sequence. Apparently, the tuning can make the ensembles more robust, leading to a better performance, and the optimization algorithm is capable of optimizing the tuned ensembles by reducing some noise or redundancy. It leads to ensembles which are less complex in terms of the actual feature counts, at the same time often showing an improved performance compared to the simple ensembles. For example, using variationist features based on CFG production rules, we generated 71 separate variables-based models corresponding

to the particular grammatical categories (NN, JJ and RB, etc., as well as NP, VP and S, etc.). Employing optimization and tuning reduced the number of models from 71 to eight, at the same time increasing the accuracy by $\approx 14\%$ compared to a simple ensemble consisting of all original 71 models. In sum, our ensemble approach seems to be a useful tool for optimizing large NLI systems.

15.2 Contributions

In this Section, we summarize the contributions of this thesis with respect to the five **research questions** presented in Section 1.3.

1. *How useful are features on different levels of linguistic modelling for the specific task of NLI?*
   [LINGUISTIC-FEATURES]

   (a) In several chapters of this thesis (i.e., Chapter 5, Chapter 6, Chapter 9 and Chapter 14), we encountered the same general pattern: Surface-based features such as word n-grams yield reasonably high accuracies, and seem to pose one of the best performing single feature types for NLI in single- as well as cross-corpus settings.[1] Thus, if one is interested in obtaining a reasonably performing baseline system with a least possible effort, using n-grams is certainly the right way to start. Our findings also suggest that using n-grams of higher length than the bi- or tri-grams usually considered in the related research, might be beneficial from the quantitative point of view – Especially if some more abstract n-gram versions are utilized (e.g., the versions POS-based or Open-Class-POS-based n-grams incorporating different amounts of Part-of-Speech information). Furthermore, combining models based on various n-gram lengths has in general a positive effect on NLI accuracy. The question what is the maximum length still worth considering is not easy to answer. It depends on different factors such as

[1] This finding is also supported by relevant related work, especially by the outcomes of the First NLI Shared Task reporting a range of different approaches to NLI (Tetreault et al., 2013).

the corpus size and the nature of the overall system. For some of our settings n-grams up to $n = 5$ (see Chapter 5), and even $n = 8$ (see Chapter 14) turned out to provide a quantitative edge. The best way is probably to explore different parameter setting for the particular tasks. However, our results suggest that n-grams with $n > 10$ are unlikely to be useful for NLI.

(b) While n-grams are of high use in terms of classification performance in NLI, it still seems worthwhile to go beyond them, and to consider some more linguistically-motivated features at different levels of linguistic modelling. Such features, modelling different linguistic units, structures and phenomena, are expected to capture interesting L1-transfer effects, which could hardly be reflected using solely n-grams. Indeed, using such linguistic features shows both, quantitative and qualitative advantages. On the one hand, extending systems based on n-grams by some morphological features, syntactic features encoding constituency and dependency information in different ways, and features reflecting language complexity, showed an increase in classification accuracy. On the other hand, employing linguistically more elaborated features, e.g., features reflecting subject realization or nominal modification patterns, enables interesting qualitative analyses capable of enhancing insight in SLA beyond what is possible using n-grams. Thus, employing features at different levels of linguistic modelling seems to be useful from the quantitative and qualitative perspectives and worth further consideration in the future work.

2. *How well do results and findings based on a broad range of features generalize across different data sets?*
[CROSS-CORPUS]

(a) We conducted a range of experiments employing a range of different corpora such as TOEFL11, ICLE, HKUST, USE, NOCE, FCE, BALC, ICNALE, and TÜTEL-NLI. On the one hand, our results show that it seems possible to obtain reasonably high cross-corpus results under certain conditions, i.e., if the training and test sets are sufficiently sim-

ilar with respect to the general structure and the genre (see Chapter 5). However, other findings in this thesis (see Chapter 6, Chapter 9 and Chapter 14), as well as related research (cf., e.g., Tetreault et al., 2013; Brooke & Hirst, 2011, 2012b) suggest that in general, obtaining robust systems showing high accuracies across arbitrary data sets remains a challenge in NLI. Even if a broad feature set is employed. Thus, this issue should be further focused on and explored in the future work.

(b) Further, our findings suggest that there are certain regularities holding for single- as well as cross-corpus experiments. In particular, the fact that lexical features such as n-grams pose one of the best performing single feature types (see Chapter 6). Also the use of certain methods seems to be equally useful for single- and cross-corpus settings – Our ensemble optimization and tuning procedures improved the results across data sets (see Part IV).

3. *How can we abstract over individual features to obtain insights into the general underlying linguistic structures reflected in NLI?*
[GENERAL-STRUCTURES]

(a) In this thesis, we explored different feature types abstracting from specific surface features to more general classes of features. For this we employed linguistic generalizations such as Part-of-Speech, phrasal categories or grammatical functions, etc. Such features are capable of capturing general underlying linguistic structures, and are first of all of interest in the context of qualitative evaluations on L1-transfer. E.g., they allowed for valuable qualitative analyses on subject realization or nominal modification patterns. Moreover, as already pointed out above such general linguistic features can also contribute to improving the classification performance in NLI, and thus are of value from the quantitative perspective as well.

(b) Further, we proposed a flexible machine learning technique – namely, label-informed features-grouping – based on hierarchical clustering (see Chapter 8). The method seems capable of generalizing from certain individual linguistic features to reasonable classes. This is done

by grouping those features together which behave structurally alike with respect to the classification label, i.e., the L1. The method is, first of all, applicable to features following the variationist sociolinguistics perspective (see Chapter 7). On the one hand, it shows quantitative gains by reducing potential data sparsity. Especially, in cases where the individual features are rather rare, but the underlying linguistic structures are common. E.g., certain verbs might be rare, but the subcategorization patters they realize might be common. Since the technique is capable of grouping verbs showing similar subcategorization patterns, the corresponding groups as features are generally less sparse than the individual verbs considered separately as features (see Chapter 9, in particular Section 9.3). On the other hand, the groups inferred by the technique provide an interesting basis for qualitative explorations on L1-transfer. In particular, the method can foster the discovery of interesting new L1-transfer candidates, and it facilitates the validation of hypotheses about L1-transfer (see Chapter 10).

4. *Can the application of variationist perspective to language data enhance an NLI system and contribute relevant SLA insight?*
   [VARIATIONIST-PERSPECTIVE]

   (a) In this thesis we proposed a variationist approach to NLI, which is at the heart of our study (see Part III, i.e., the Chapters 7–11). We explored it from quantitative and qualitative point of view.

   (b) The variationist perspective is based on the notion of a linguistic variable, which can be realized by a set of variants. We discussed this notion in detail and proposed a revised version, which is more flexible and which we consider better suitable in the given context. Further, we proposed a taxonomy of linguistic variables useful in the context of tasks such as NLI. Moreover, we investigated employing the label-informed features-grouping method proposed in Chapter 8, for generating more abstract linguistic variables.

   (c) We employed different variationist features focusing at the syntax level,

and showed that the approach is capable of advancing the classification accuracy of NLI systems (see Chapter 9 and Chapter 14).

(d) Our qualitative explorations utilizing variationist features encoding subject realization in general, subject pronoun realization in particular, as well as nominal modification, provided instructive outcomes. We exemplified a possible way of discovering interesting L1-transfer candidates as well as testing and validating new and existing hypotheses about L1-transfer (see Chapter 10).

5. *How can we optimize large models incorporating a broad range of features?* [MODEL-OPTIMIZATION]

(a) We showed that the proposed label-informed feature grouping technique (see Chapter 8) is capable of optimizing feature sets by grouping features together behaving alike with respect to the classification label (L1). The method can make the models more compact, and thus more efficient by reducing the feature space.

(b) We explored combining a broad range of feature types using meta-classifier techniques. In particular, we employed a probability-based ensemble, combining the probability estimates for the different L1, yielded by individual logistic regression classifiers based on different feature types. This method outperformed combining the different features directly in a single vector, which confirms the findings from the previous research.

(c) We proposed an ensemble optimization method, utilizing a linear model selection algorithm, capable of selecting reasonable model subsets. Some of the generated sub-systems outperformed the overall system.

(d) We suggested an ensemble tuning procedure, which is based on including all of the available features in a condensed, abstract form into the ensemble, showing reasonable quantitative advantages.

(e) Combining tuning and optimization in sequence, seems to pose an appropriate method for optimizing large models, and further advancing the performance.

(f) Applying our ensemble approach to a wide range of surface-based and linguistically-motivated features, showed an accuracy of 85.5%, outperforming the best results published so far for the standard TOEFL11 data setup.

Finally, *"Is there choice in non-native voice"*, or is there no real choice? Is the choice maybe predetermined by the L1? Since the accuracies yielded by our system – including the variationist models, which are designed to explicitly encode the choices –, are well above chance, the choice definitely seems to be *influenced* by the L1 of the learners. However, there are many other factors, such as proficiency, social status, age or gender potentially influencing the choices and preferences. Unfortunately, it is not possible to control for all of them with the given data. Thus, further dedicated investigations are required for a better understanding of the effect and the interplay of the different factors on the choices made by the learners. One could also argue that learners have an even *"greater choice"* than the natives – They show regular and various irregular realizations, and thus can have an extended set of options compared to the natives, generally realizing regular forms. In sum, our study contributes a piece to the overall puzzle behind this question, which, in our opinion, is worth further explorations.

15.3 Limitations and Outlook

We explored several feature types at the lexical level, and at the levels of morphology and syntax. However, we did not employ any explicit semantic or pragmatic modelling. For example, we did not consider the discourse structure or the realized speech acts, etc. in our modelling. Nevertheless, corresponding features have the potential of providing further valuable and instructive findings. Engineering appropriate features and exploring them from variationist perspective might be worthwhile considering in terms of the future work.

The employed off-the-shelf NLP tools and models, have not been explicitly designed for the processing of learner language. However, this language variety shows certain peculiarities such as incomplete sentences or missing punctuation, etc. This poses a challenge for standard NLP tools, resulting in usually increased

error rates in tagging and parsing, etc. Thus, generating and utilizing models adjusted to learner data might further improve the results.

For our study, we employed several corpora varying in size and quality. For most of the explorations we focused on a corpus, we consider most suitable for our explorations, namely the TOEFL11, designed specifically for the task of NLI. However, employing other reasonably sized corpora including a wide range of different L1s, and controlling for various relevant parameters such as prompt, proficiency, age and gender, etc., would allow for more fine-grained analyses, and further assessment of the advantages and limitations of the approach.

In our qualitative explorations, we consciously avoided any strong conclusions about L1-transfer, speaking about L1-transfer *candidates* instead. We believe that any findings should be verified using a range of appropriate data sets, applying different suitable methods, before making any strong conclusions.

Overcoming these limitation in terms of the future work will certainly lead to new valuable findings.

Zusammenfassung

Ist es möglich die Muttersprache eines Autors anhand eines nicht-muttersprachlichen Textes zu erkennen? Lässt sich diese Aufgabe vollständig automatisiert bewältigen? Ein hohes Interesse an Antworten auf diese Fragen führte zu Beginn dieses Jahrhunderts zur Entstehung des neuen Forschungsfeldes *automatische Muttersprachenerkennung* (engl. *Native Language Identification*, kurz *NLI*). Sprachliche Daten als Grundlage auf der einen Seite, sowie die Anforderung, automatisch ein bestimmtes Merkmal des Autors anhand dieser Daten abzuleiten auf der anderen Seite, situiert die gegebene Aufgabe in der Schnittmenge zwischen Linguistik und Informatik, bzw. in der Computerlinguistik als der Disziplin, welche die beiden oben genannten in sich vereint.

Die vorliegende Arbeit nimmt sich einiger relevanter Forschungsfragen im Bereich von NLI an: Was ist die Rolle von Oberflächenmerkmalen und wie wichtig sind abstraktere linguistische Eigenschaften? Wie ist eine Vielzahl von Merkmalen zu einem System zusammenzuführen, und wie können die resultierenden komplexen Modelle optimiert werden? Inwiefern können die auf bestimmten Daten gewonnenen Erkenntnisse von allgemeiner Gültigkeit sein? Kann die Betrachtung der gegebenen Aufgabe im Kontext der variationslinguistischen Theorie gewinnbringend sein?

Um Antworten auf diese Fragen zu finden, haben wir im Rahmen dieser Dissertation eine Reihe von quantitativen sowie qualitativen Untersuchungen unter Verwendung von statistischen Verfahren aus dem Bereich des maschinellen Lernens durchgeführt.

Insbesondere haben wir gezeigt, dass Oberflächenmerkmale von hoher Bedeutung für die Genauigkeit von NLI-Systemen sind, die Verwendung von abstrakteren linguistischen Merkmalen jedoch zu weiteren Verbesserungen führt. Die

Letzteren sind auch von besonderem Interesse im Zusammenhang mit der qualitativen Analyse und Interpretation der Ergebnisse, da sie bestimmte sprachliche Strukturen reflektieren können, welche durch die Verwendung von Oberflächenmerkmalen allein kaum zugänglich wären.

Ferner haben wir als Datengrundlage eine Reihe von Korpora eingesetzt und gezeigt, dass Erkenntnisse zu NLI, die auf der Basis eines bestimmten Korpus gewonnen worden sind, sich durchaus auch auf andere Daten übertragen lassen, sofern diese von ähnlicher Natur und Struktur sind. Robuste NLI-Systeme, die gute Ergebnisse für Daten unterschiedlicher Beschaffenheit liefern können, stellen jedoch weiterhin eine große Herausforderung im gegebenen Forschungsbereich dar.

Des Weiteren haben wir im Rahmen der vorliegenden Arbeit einen variationslinguistischen Ansatz für NLI entwickelt und gezeigt, dass solch eine Perspektive sowohl zur Verbesserung der quantitativen Ergebnisse als auch zu interessanten qualitativen Erkenntnissen führen kann. So zeigen unsere Ergebnisse wertvolle Einsichten zu Mustern im Zweitspracherwerb im Zusammenhang mit der Subjektrealisierung sowie Modifikation von Nominalphrasen. Basierend auf der variationslinguistischen Perspektive, haben wir ein statistisches Verfahren zur Extraktion von abstrakteren linguistischen Merkmalen entwickelt, deren Verwendung sowohl zur Verbesserung der Genauigkeit von NLI-Systemen führen kann, als auch weiterführende linguistische Analysen ermöglicht. Unter Verwendung unseres Ansatzes lassen sich sowohl existierende Hypothesen aus dem Bereich des Zweitspracherwerbs überprüfen, als auch neue Hypothesen bilden und validieren.

Schließlich haben wir eine Reihe von sprachlichen Merkmalen unterschiedlicher Natur zu einem komplexen NLI-System kombiniert, und gezeigt wie solche Modelle optimiert werden können. Unser System erzielte eine Vorhersagegenauigkeit von 85.5% bei der Unterscheidung zwischen 11 verschiedenen Muttersprachen – Das beste bis dato veröffentlichte Ergebnis, das in einem standardisierten Testverfahren im gegebenen Forschungsbereich erzielt worden ist.

Insgesamt zeigt die vorliegende Arbeit, wie linguistische Erkenntnisse zur Verbesserung von Technologie beitragen können, und wie, im Gegenzug, die Technologie die Gewinnung von neuen Erkenntnissen in der linguistischen Theorie begünstigen kann – Ein viel versprechendes und fruchtbares Wechselspiel.

Bibliography

Aharodnik, K., M. Chang, A. Feldman & J. Hana (2013). Automatic Identification of Learners' Language Background Based on Their Writing in Czech. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Nagoya, Japan: Asian Federation of Natural Language Processing, pp. 1428–1436.

Alpaydin, E. (2004). *Introduction to Machine Learning*. Cambridge / London: The MIT Press.

Amaral, L. & D. Meurers (2008). 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.

Argamon, S., M. Koppel, J. W. Pennebaker & J. Schler (2009). Automatically Profiling the Author of an Anonymous Text. In *Communications of the ACM*. vol. 52, pp. 119–123.

Aristotle (1933). *The Metaphysics, Books 1-9*. Cambridge: Harvard University Press. Translated by H. Tredennick.

Axelsson, M. W. (2000). USE – The Uppsala Student English Corpus: An instrument for needs analysis. *ICAME Journal* 24, 155–157. URL `http://icame.uib.no/ij24/use.pdf`.

Axelsson, M. W. (2003). *Manual: The Uppsala Student English Corpus (USE)*. Uppsala University, Department of English, Sweden. URL `http://www.`

engelska.uu.se/Research/English_Language/Research_Areas/Electronic_Resource_Projects/USE-Corpus.

Baayen, R. H., R. Piepenbrock & L. Gulikers (1995). The CELEX Lexical Databases. CDROM. URL http://www.ldc.upenn.edu/Catalog/readme_files/celex.readme.html.

Bally, C. (1944). *Linquistique générale et linguistique française*. Berne: Francke.

Baroni, M. & S. Bernardini (2006). A New Approach to the Study of Translationese: Machinelearning the Difference between Original and Translated Text. *Literary and Linguistic Computing* 21(3), 259–274.

Bestgen, Y., S. Granger & J. Thewissen (2012). Error Patterns and Automatic L1 Identification. In S. Jarvis & S. A. Crossley (eds.), *Approaching Language Transfer through Text Classification: Explorations in the Detection-based Approach*, Bristol: Multilingual Matters, pp. 127–153.

Björkelund, A., B. Bohnet, L. Hafdell & P. Nugues (2010). A high-performance syntactic and semantic dependency parser. In *Demonstration Volume of the 23rd International Conference on Computational Linguistics (COLING 2010)*. Beijing, China, pp. 23–27. URL https://code.google.com/p/mate-tools/.

Blanchard, D., J. Tetreault, D. Higgins, A. Cahill & M. Chodorow (2013). *TOEFL11: A Corpus of Non-Native English*. Tech. rep., Educational Testing Service.

Bohnet, B. (2010). Top Accuracy and Fast Dependency Parsing is not a Contradiction. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*. Beijing, China, pp. 89–97.

Boyd, A., M. Dickinson & D. Meurers (2008). On Detecting Errors in Dependency Treebanks. *Research on Language and Computation* 6(2), 113–137.

Brooke, J. & G. Hirst (2011). Native Language Detection with 'Cheap' Learner Corpora. Presented at the *Learner Corpus Research (LCR 2011)*. Louvain-la-Neuve.

Brooke, J. & G. Hirst (2012a). Measuring Interlanguage: Native Language Identification with L1-influence Metrics. In *Proceedings of the 8th ELRA Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul, pp. 779–784.

Brooke, J. & G. Hirst (2012b). Robust, Lexicalized Native Language Identification. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*. Mumbai, India, pp. 391–408.

Brooke, J. & G. Hirst (2013). Using Other Learner Corpora in the 2013 NLI Shared Task. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-8) at NAACL-HLT 2013*. Atlanta, GA, pp. 188–196.

Brunner, T. (2015). The structure of the noun phrase in Singaporean and Kenyan English: a corpus-based study. Ph.D. thesis, Universität Regensburg, Germany.

Brysbaert, M., M. Buchmeier, M. Conrad, A. Jacobs, J. Bölte & A. Böhl (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology* 58, 412–424.

Burton, K., A. Java & I. Soboroff (2009). The ICWSM 2009 Spinn3r Dataset. In *Proceedings of the Third Annual Conference on Weblogs and Social Media (ICWSM)*. San Jose, CA.

Butterworth, R., G. Piatetsky-Shapiro & D. A. Simovici (2005). On Feature Selection through Clustering. In *Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'2005)*. pp. 581–584.

Bykh, S. (2011). "Hat Wiederholung Stil?" – Eine computerunterstützte stilometrische Untersuchung zu repetitiven Wortsequenzen. Master's thesis, University of Tübingen, Germany.

Bykh, S. & D. Meurers (2012). Native Language Identification Using Recurring N-grams – Investigating Abstraction and Domain Dependence. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*. Mumbai, India, pp. 425–440.

Bykh, S. & D. Meurers (2014). Exploring Syntactic Features for Native Language Identification: A Variationist Perspective on Feature Encoding and Ensemble Optimization. In *Proceedings of the 25th International Conference on Computational Linguistics (2014)*. Dublin, Ireland, pp. 1962–1973.

Bykh, S. & D. Meurers (2016). Advancing Linguistic Features and Insights by Label-informed Feature Grouping: An Exploration in the Context of Native Language Identification. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*. Osaka, Japan, pp. 739–749.

Bykh, S., S. Vajjala, J. Krivanek & D. Meurers (2013). Combining Shallow and Linguistically Motivated Features in Native Language Identification. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-8) at NAACL-HLT 2013*. Atlanta, GA, pp. 197–206.

Callies, M. & K. Szczesniak (2008). Argument realization, information status and syntactic weight – A learner-corpus study of the dative alternation. In M. Walter & P. Grommes (eds.), *Fortgeschrittene Lernervarietäten*, Tübingen: Niemeyer, pp. 165–187.

Callies, M. & E. Zaytseva (2011). The Corpus of Academic Learner English (CALE): A new resource for the study of lexico-grammatical variation in advanced learner varieties. In H. Hedeland, T. Schmidt & K. Wörner (eds.), *Multilingual Resources and Multilingual Applications*, Hamburg Working Papers in Multilingualism, B 96, pp. 51–56.

Cedergren, H. J. & D. Sankoff (1974). Variable rules: Performance as a statistical reflection of competence. *Language* 50(2), 333–355.

Chang, C.-C. & C.-J. Lin (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2, 27:1–27:27. URL `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Choroleeva, K. (2009). Language Transfer: Types of Linguistic Errors Committed by Francophones Learning English as a Second Foreign Language. *Humanising Language Teaching* 11(5).

Cimino, A., F. Dell'Orletta, G. Venturi & S. Montemagni (2013). Linguistic Profiling based on General–purpose Features and Native Language Identification. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-8) at NAACL-HLT 2013*. Atlanta, GA, pp. 207–215.

Corder, S. P. (1967). The Significance of Learner's Errors. *International Rewiew of Applied Linguistics in Language Teaching* 5(4), 161–170.

Coulmas, F. (ed.) (1997). *The Handbook of Sociolinguistics*. Massachusetts: Blackwell.

Coupland (2007). *Style: Language Variation and Identity*. Cambridge: Cambridge University Press.

Crossley, S. A. & D. S. McNamara (2012). Detecting the First Language of Second Language Writers Using Automated Indices of Cohesion, Lexical Sophistication, Syntactic Complexity and Conceptual Knowledge. In S. Jarvis & S. A. Crossley (eds.), *Approaching Language Transfer through Text Classification: Explorations in the Detection-based Approach*, Bristol: Multilingual Matters, pp. 106–126.

Daelemans, W., J. Zavrel, K. van der Sloot & A. van den Bosch (2007). *TiMBL: Tilburg Memory-Based Learner Reference Guide, ILK Technical Report ILK 07-03*. Induction of Linguistic Knowledge Research Group Department of Communication and Information Sciences, Tilburg University, Tilburg, The Netherlands. URL `http://ilk.uvt.nl/downloads/pub/papers/ilk.0703.pdf`.

de Marneffe, M.-C., B. MacCartney & C. Manning (2006). Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*. Genoa, Italy, pp. 449–454.

Dechert, H. & M. Raupach (eds.) (1989). *Transfer in Language Production*. Norwood: Ablex.

Dickinson, M., C. Brew & D. Meurers (2013). *Language and Computers*. Wiley-Blackwell.

Dickinson, M. & W. D. Meurers (2003). Detecting Errors in Part-of-Speech Annotation. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03)*. Budapest, Hungary, pp. 107–114.

Dickinson, M. & W. D. Meurers (2005). Detecting Errors in Discontinuous Structural Annotation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*. pp. 322–329.

Dines, E. (1980). Variation in discourse - and stuff like that. *Language in Society* 9, 13–32.

Domínguez, L. (2013). *Understanding Interfaces: Second Language Acquisition and First Language Attrition of Spanish Subject Realization and Word Order Variation*, vol. 55 of *Language Acquisition and Language Disorders*. Amsterdam: John Benjamins.

Doughty, C. & M. Long (eds.) (2003a). *The Handbook of Second Language Acquisition*. Blackwell Handbooks in Linguistics. Oxford: Blackwell Publishing.

Doughty, C. J. & M. Long (2003b). The Scope of Inquiry and Goals of SLA. In C. Doughty & M. Long (eds.), *The Handbook of Second Language Acquisition*, Oxford: Blackwell Publishing, pp. 3–16.

Duden (2003). *Deutsches Universalwörterbuch*. Mannheim: Dudenverlag.

Díaz Negrillo, A. (2007). A Fine-Grained Error Tagger for Learner Corpora. Ph.D. thesis, University of Jaén, Spain.

Díaz Negrillo, A. (2009). *EARS: A User's Manual*. Munich: LINCOM Academic Reference Books.

Eisenberg, P. (1994). *Grundriß der deutschen Grammatik*. Stuttgart: Metzler.

Ellis, N. (2013). Construction Grammar and Second Language Acquisition. In T. Hoffmann & G. Trousdale (eds.), *The Oxford Handbook of Construction Grammar*, Oxford: Oxford University Press, pp. 365–378.

Estival, D., T. Gaustad, S. Pham, W. Radford & B. Hutchinson (2007). Author profiling for English emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*. pp. 263–272.

Fan, R., K. Chang, C. Hsieh, X. Wang & C. Lin (2008). LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research* 9, 1871–1874. URL `http://www.csie.ntu.edu.tw/~cjlin/liblinear`.

Gass, S. (1996). Second language acquisition and linguistic theory: the role of language transfer. In W. Ritchie & T. K. Bhatia (eds.), *Handbook of Second Language Acquisition*, San Diego: Academic Press, pp. 317–345.

Gass, S. & L. Selinker (eds.) (1983). *Language Transfer in Language Learning*. Rowley: Newbury House.

Gass, S. M., J. Behney & L. Plonsky (2013). *Second Language Acquisition: An Introductory Course*. New York / London: Routledge.

Gass, S. M. & L. Selinker (eds.) (1992). *Language Transfer in Language Learning*. Amsterdam: John Benjamins.

GBN (2016). Google Books Ngrams. URL `https://books.google.com/ngrams`. Resource accessed 2016.

Gebre, B. G., M. Zampieri, P. Wittenburg & T. Heskes (2013). Improving Native Language Identification with TF-IDF Weighting. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-8) at NAACL-HLT 2013*. Atlanta, GA, pp. 216–223.

Geertzen, J., T. Alexopoulou, R. Baker, H. Hendriks, S. Jiang, A. Korhonen & E. E. First (2013). *The EF Cambridge Open Language Database (EFCAMDAT), user manual, part I*. written production.

Geertzen, J., T. Alexopoulou & A. Korhonen (2014). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In *Proceedings of the 31st Second Language Research Forum (SLRF)*. Somerville: Cascadilla Press, pp. 240–254.

Geeslin, K. L. & A. Y. Long (2014). *Sociolinguistics and Second Language Acquisition*. New York: Routledge.

Giacalone Ramat, A. (ed.) (2003). *Typology and Second Language Acquisition*. Berlin: De Gruyter Mouton.

Golcher, F. & M. Reznicek (2011). Stylometry and the interplay of topic and L1 in the different annotation layers in the FALKO corpus. In *Proceedings of Quantitative Investigations in Theoretical Linguistics 4*. Berlin, Germany, pp. 29–34.

Graesser, A. C., D. S. McNamara & J. M. Kulikowich (2012). Coh-Metrix: Providing Multilevel Analyses of Text Characteristics. *Educational Researcher* 40(5), 223–234.

Granger, S., E. Dagneaux & F. Meunier (2002a). *International Corpus of Learner English*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Granger, S., E. Dagneaux, F. Meunier & M. Paquot (2009). *International Corpus of Learner English, Version 2*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Granger, S., J. Hung & S. Petch-Tyson (eds.) (2002b). *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins.

Greenberg, J. H. (1971). *Language, culture, and communication*. Stanford: Stanford University Press.

Gundel, J. K. & E. E. Tarone (1992). Language Transfer and the Acquisition of Pronouns. In S. M. Gass & S. Selinker (eds.), *Language Transfer in Language Learning*, Amsterdam: John Benjamins, pp. 87–100.

Guo, Y. & G. H. Beckett (2007). The Hegemony of English as a Global Language: Reclaiming Local Knowledge and Culture in China. *Convergence* 40(1-2), 117–132.

Hacohen, A. & J. Schaeffer (2007). Subject realization in early Hebrew/English bilingual acquisition: The role of crosslinguistic influence. *Bilingualism: Language and Cognition* 10(3), 333–344.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann & I. H. Witten (2009). The WEKA Data Mining Software: An Update. In *The SIGKDD Explorations*. vol. 11, pp. 10–18.

Han, B., P. Cook & T. Baldwin (2012). Geolocation Prediction in social media data by finding location indicative words. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*. Mumbai, India, pp. 1045–1061.

Henderson, J., G. Zarrella, C. Pfeifer & J. D. Burger (2013). Discriminating Non-Native English with 350 Words. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-8) at NAACL-HLT 2013*. Atlanta, GA, pp. 101–110.

Hieble, J. (1957). Compound Words in German. *The German Quarterly* 30(3), 187–190.

Hirschmann, H., A. Lüdeling, I. Rehbein, M. Reznicek & A. Zeldes (2013). Underuse of Syntactic Categories in Falko. A Case Study on Modification. In S. Granger, G. Gilquin & F. Meunier (eds.), *20 years of learner corpus research. Looking back, Moving ahead. Corpora and Language in Use – Proceedings 1*, Louvain la Neuve: Presses universitaires de Louvain, pp. 223–234.

Holmes, D. I. (1994). Authorship Attribution. *Computers and the Humanities* 28, 87–106.

Hoover, D. L. (2002). Frequent word sequences and statistical stylistics. *Literary and Linguistic Computing* 17(2), 157–180.

Huddleston, R. & G. K. Pullum (2002). *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Ionescu, R. T., M. Popescu & A. Cahill (2014). Can characters reveal your native language? A language-independent approach to native language identification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar, pp. 1363–1373.

Ishikawa, S. (2011). A new horizon in learner corpus studies: The aim of the ICNALE projects. In G. Weir, S. Ishikawa & K. Poonpon (eds.), *Corpora and language technologies in teaching, learning and research*, Glasgow: University of Strathclyde Publishing, pp. 3–11. `http://language.sakura.ne.jp/icnale/index.html`.

Jarvis, S., Y. Bestgen, S. A. Crossley, S. Granger, M. Paquot, J. Thewissen & D. S. McNamara (2012a). The Comparative and Combined Contributions of n-Grams, Coh-Metrix Indices and Error Types in the L1 Classification of Learner Texts. In S. Jarvis & S. A. Crossley (eds.), *Approaching Language Transfer through Text Classification: Explorations in the Detection-based Approach*, Bristol: Multilingual Matters, pp. 154–177.

Jarvis, S., Y. Bestgen & S. Pepper (2013). Maximizing Classification Accuracy in Native Language Identification. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-8) at NAACL-HLT 2013*. Atlanta, GA, pp. 111–118.

Jarvis, S., G. Castañeda-Jiménez & R. Nielsen (2012b). Detecting L2 Writers' L1s on the Basis of Their Lexical Styles. In S. Jarvis & S. A. Crossley (eds.), *Approaching Language Transfer through Text Classification: Explorations in the Detection-based Approach*, Bristol: Multilingual Matters, pp. 34–70.

Jarvis, S. & S. A. Crossley (eds.) (2012). *Approaching Language Transfer through Text Classification: Explorations in the Detection-based Approach*, vol. 64 of *Second Language Acquisition*. Bristol: Multilingual Matters.

Jarvis, S. & M. Paquot (2012). Exploring the Role of n-Grams in L1-Identification. In S. Jarvis & S. A. Crossley (eds.), *Approaching Language*

Transfer through Text Classification: Explorations in the Detection-based Approach, Bristol: Multilingual Matters, pp. 71–105.

Jiang, X., Y. Guo, J. Geertzen, D. Alexopoulou, L. Sun & A. Korhonen (2014). Native Language Identification Using Large, Longitudinal Data. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. pp. 3309–3312.

Johansson, S. (2007). *Seeing through Multilingual Corpora: On the use of corpora in contrastive studies*. Amsterdam: John Benjamins.

Jurafsky, D. & J. H. Martin (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River: Prentice Hall.

Kanehira, Y. (2003). An Analysis of L2 Learners' Use of Noun Modification in Written Japanese. In *Proceedings of the 2003 Conference of the Australian Linguistic Society*. Newcastle, Australia, pp. 1–15.

Klein, D. & C. D. Manning (2002). Fast Exact Inference with a Factored Model for Natural Language Parsing. In *Advances in Neural Information Processing Systems 15 (NIPS 2002)*. Cambridge / London: The MIT Press. URL `http://books.nips.cc/papers/files/nips15/CS01.pdf`.

Koppel, M., J. Schler & K. Zigdon (2005). Determining an author's native language by mining a text for errors. In *Proceedings of the 11th ACM SIGKDD international conference on Knowledge discovery in data mining (KDD)*. New York, NY, pp. 624–628.

Krashen, S. D. (1982). *Principles and Practice in Second Language Acquisition*. Oxford: Pergamon Press.

Krier, C., D. François, F. Rossi & M. Verleysen (2007). Feature clustering and mutual information for the selection of variables in spectral data. In *Proceedings of the ESANN'2007*. Bruges, Belgium, pp. 157–162.

Krivanek, J. (2012). Investigating Syntactic Alternations as Characteristic Features of Learner Language. Master's thesis, University of Tübingen, Germany.

Kubat, M. (2015). *An introduction to machine learning*. Cham / Heidelberg: Springer.

Kuperman, V., H. Stadthagen-Gonzalez & M. Brysbaert (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods* 44(4), 978–990.

Labov, W. (1969). Contraction, deletion, and inherent variability of the English copula. *Language* 45(4), 715–762.

Labov, W. (1972). *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Lado, R. (1957). *Linguistics Across Cultures: Applied Linguistics for Language Teachers*. Ann Arbor: University of Michigan Press.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. Chicago: University of Chicago Press.

Levy, R. & G. Andrew (2006). Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *5th International Conference on Language Resources and Evaluation (LREC)*. Genoa, Italy.

Long, M. H. & C. J. Sato (1984). Methodological issues in interlanguage studies: an interactionist perspective. In A. Davies, C. Criper & A. Howatt (eds.), *Interlanguage*, Edinburgh: Edinburgh University Press, pp. 253–279.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* 15(4), 474–496.

Lu, X. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Languages Journal* 96(2), 190–208.

Lüdeling, A. (2011). Corpora in Linguistics: Sampling and Annotation. In K. Grandin (ed.), *[Nobel Symposium 147] Going Digital: Evolutionary and Revolutionary Aspects of Digitization*. New York: Science History Publications, pp. 220–243.

Lynum, A. (2013). Native Language Identification Using Large Scale Lexical Features. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-8) at NAACL-HLT 2013*. Atlanta, GA, pp. 266–269.

Malmasi, S. & A. Cahill (2015). Measuring Feature Diversity in Native Language Identification. In *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-10) at NAACL-HLT 2015*. Denver, Colorado, pp. 49–55.

Malmasi, S. & M. Dras (2014a). Arabic Native Language Identification. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP)*. Doha, Qatar, pp. 180–186.

Malmasi, S. & M. Dras (2014b). Chinese Native Language Identification. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Gothenburg, Sweden, pp. 95–99.

Malmasi, S. & M. Dras (2014c). Finnish Native Language Identification. In *Proceedings of the Australasian Language Technology Association Workshop 2014*. Brisbane, Australia, pp. 139–144.

Malmasi, S. & M. Dras (2014d). Language Transfer Hypotheses with Linear SVM Weights. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar, pp. 1385–1390.

Malmasi, S. & M. Dras (2015a). Large-scale Native Language Identification with Cross-Corpus Evaluation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL 2015)*. Denver, Colorado, pp. 1403–1409.

Malmasi, S. & M. Dras (2015b). Multilingual native language identification. *Natural Language Engineering* 1(1), 1–87.

Malmasi, S., M. Dras & I. Temnikova (2015a). Norwegian Native Language Identification. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*. Hissar, Bulgaria, pp. 404–412.

Malmasi, S., J. Tetreault & M. Dras (2015b). Oracle and Human Baselines for Native Language Identification. In *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-10) at NAACL-HLT 2015*. Denver, Colorado, pp. 172–178.

Manning, C. & H. Schütze (1999). *Foundations of statistical Natural Language Processing*. Cambridge: The MIT Press.

Meurers, D., J. Krivanek & S. Bykh (2014). On the Automatic Analysis of Learner Corpora: Native Language Identification as Experimental Testbed of Language Modeling between Surface Features and Linguistic Abstraction. In *Diachrony and Synchrony in English Corpus Studies*. Frankfurt am Main: Peter Lang, pp. 285–314.

MFGW (2016). The Most Frequent German Words. LSA, University of Michigan. URL `https://www.lsa.umich.edu/german/hmr/Vokabeln/frequent_words.html`. Resource accessed 2016.

Milton, J. C. P. & N. Chowdhury (1994). Tagging the interlanguage of Chinese learners of English. In *Proceedings joint seminar on corpus linguistics and lexicology*. Guangzhou and Hong Kong, Language Centre, HKUST, pp. 127–143. URL `http://hdl.handle.net/1783.1/1087`.

Mosteller, F. & D. Wallace (1964). *Inference and Disputed Authorship: The Federalist*. Reading: Addison-Wesley.

Nicolai, G. & G. Kondrak (2014). Does the Phonology of L1 Show Up in L2 Texts? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Baltimore, Maryland, pp. 854–859.

Odlin, T. (1989). *Language Transfer: Cross-linguistic influence in language learning*. Cambridge: Cambridge University Press.

Odlin, T. (2003). Cross-linguistic Influence. In C. J. Doughty & M. H. Long (eds.), *Handbook on Second Language Acquisition*, Oxford: Blackwell Publishing, pp. 436–486.

Oliva, M. A. A. & M. J. Serrano (2013). *Style in Syntax: Investigating variation in Spanish pronoun subjects*. Bern: Peter Lang.

Ortega, L. (2009). *Understanding Second Language Acquisition*. London: Hodder Education.

Park, C. H. (2013). A Feature Selection Method Using Hierarchical Clustering. In R. Prasath & T. Kathirvalavakumar (eds.), *Proceedings of the International Conference on Mining Intelligence and Knowledge Exploration (MIKE 2013), LNAI 8284*. Virudhunagar, Madurai, India: Springer International Publishing Switzerland, pp. 1–6.

Pate, J. & D. Meurers (2007). Refining Syntactic Categories Using Local Contexts – Experiments in Unlexicalized PCFG Parsing. In *Proceedings of the Sixth Workshop on Treebanks and Linguistic Theories (TLT 2007)*. Bergen, Norway.

Petrov, S. & D. Klein (2007). Improved Inference for Unlexicalized Parsing. In *Proceedings of the Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007)*. Rochester, NY, pp. 404–411.

Platt, J. C. (1998). *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*. Tech. Rep. MSR-TR-98-14, Microsoft Research.

Pollard, C. & I. A. Sag (1994). *Head-Driven Phrase Structure Grammar*. Chicago: The University of Chicago Press.

Randall, M. & N. Groom (2009). The BUiD Arab Learner Corpus: a resource for studying the acquisition of L2 English spelling. In *Proceedings of the Corpus Linguistics Conference (CL)*. Liverpool, UK.

Richards, J. C. (1971). A non-contrastive approach to error analysis. *ELT Journal* 25(3), 204–219.

Richards, J. C. (1974). *Error analysis: perspectives on second language acquisition*. London: Longman.

Robertson, D. (2000). Variablility in the use of the English article system by Chinese learners of English. *Second Language Research* 16(2), 135–172.

Sankoff, D. & P. Thibault (1981). Weak complementarity: Tense and aspect in Montreal French. In B. B. Johns & D. R. Strong (eds.), *Syntactic Change*, Ann Arbor: University of Michigan Press, pp. 205–216.

Santorini, B. (1990). *Part-of-speech Tagging Guidelines for the Penn Treebank, 3rd Revision, 2nd Printing*. Tech. rep., Department of Computer Science, University of Pennsylvania. URL `ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz`.

Scott, S. & S. Matwin (1999). Feature Engineering for Text Classification. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML 1999)*. San Francisco, CA: Morgan Kaufmann Publishers Inc., pp. 379–388.

Selinker, L. (1969). Language Transfer. *General Linguistics* 9, 67–92.

Selinker, L. & U. Lakshmanan (1992). Language transfer and fossilization: The "Multiple Effects Principle". In S. M. Gass & S. Selinker (eds.), *Language Transfer in Language Learning*, Amsterdam: John Benjamins, pp. 197–216.

Shi, W. (2015). Types of Chinese Negative Transfer to English Learning and the Countermeasures. *Theory and Practice in Language Studies* 5(6), 1226–1232.

Stanojević, M. (2009). Cognitive synonymy: a general overview. *Facta Universitatis, Linguistics and Literature series* 7(2), 193–200.

Stockwell, R., J. Bowen & J. Martin (1965). *The Grammatical Structures of English and Spanish*. Chicago: University of Chicago Press.

Stringer, D. (2008). What Else Transfers? In R. S. et al. (ed.), *Proceedings of the 9th Generative Approaches to Second Language Acquisition Conference (GASLA 2007)*. Somerville, MA: Cascadilla Proceedings Project, pp. 233–241.

Swanson, B. & E. Charniak (2013). Extracting the Native Language Signal for Second Language Acquisition. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL 2013)*. Atlanta, Georgia, pp. 85–94.

Swanson, B. & E. Charniak (2014). Data Driven Language Transfer Hypotheses. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguisticsi (EACL-14)*. Gothenburg, Sweden, pp. 169–173.

Tagliamonte, S. A. (2012). *Variationist Sociolinguistics: Change, Observation, Interpretation*. John Wiley & Sons.

Tetreault, J., D. Blanchard & A. Cahill (2013). A Report on the First Native Language Identification Shared Task. In *Proceedings of the 8th Workshop on Building Educational Applications Using NLP (BEA-8) at NAACL-HLT 2013*. Atlanta, GA, pp. 48–57.

Tetreault, J., D. Blanchard, A. Cahill & M. Chodorow (2012). Native Tongues, Lost and Found: Resources and Empirical Evaluations in Native Language Identification. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*. Mumbai, India, pp. 2585–2602.

Tomokiyo, L. M. & R. Jones (2001). You're Not From Round Here, Are You? Naive Bayes Detection of Non-native Utterance Text. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*. Pittsburgh, PA, pp. 239–246.

Tono, Y. (2004). Multiple comparisons of IL, L1 and TL corpora: the case of L2 acquisition of verb subcategorization patterns by Japanese learners of English. In G. Aston, S. Bernardini & D. Stewart (eds.), *Corpora and Language Learners*, Amsterdam: John Benjamins, pp. 45–66.

Tschirner, E. & R. Jones (2006). *A Frequency Dictionary of German: Core Vocabulary for Learners*. Routledge Frequency Dictionaries. London: Routledge.

Tsur, O. & A. Rappoport (2007). Using Classifier Features for Studying the Effect of Native Language on the Choice of Written Second Language Words. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition (CACLA '07)*. Prague, Czech Republic, pp. 9–16.

Vajjala, S. & D. Meurers (2012). On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. In J. Tetreault, J. Burstein & C. Leacock (eds.), *In Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-7) at NAACL-HLT 2012*. Montréal, Canada, pp. 163–173.

Voyles, J. B. (1967). German Noun and Adjective Compounds. *Language Learning* 17(1–2), 9–19.

Vyatkina, N., H. Hirschmann & F. Golcher (2015). Syntactic modification at early stages of L2 German writing development: A longitudinal learner corpus study. *Journal of Second Language Writing* 29, 28–50.

Wang, X. (2009). Exploring the Negative Transfer on English Learning. *Asian Social Science* 5(7), 138–143.

Weber, H. (1971). *Das erweiterte Adjektiv- und Partizipialattribut im Deutschen*. München: Hueber.

Weber, H. (2012). Relationelle, synthetische oder onomasiologische Grammatik? In F. Grucza (ed.), *Vielheit und Einheit der Germanistik weltweit*, Frankfurt a. M. / Berlin: Lang, vol. 15 of *Publikationen der Internationalen Vereinigung für Germanistik (IVG)*, pp. 273–277.

Weber, H. (2014). Erweitertes Partizipialattribut und Relativsatz: Ein Fall von syntaktischer Synonymie. In E. Żebrowska, M. Jaworska & D. Steinhoff (eds.), *Materiality and Mediality of Linguistic Communication: Proceedings of the 47th Linguistics Colloquium*, Olsztyn, Poland: Lang, vol. 32 of *Linguistik International*, pp. 467–477.

Weinreich, U. (1953). *Languages in Contact*. The Hague: Mouton.

Witten, I. H., E. Frank & M. A. Hall (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Amsterdam / Boston: Morgan Kaufmann.

Wong, S.-M. J. & M. Dras (2009). Contrastive analysis and native language identification. In *Australasian Language Technology Association Workshop 2009*. Sydney, Australia, pp. 53–61.

Wong, S.-M. J. & M. Dras (2011). Exploiting Parse Structures for Native Language Identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK., pp. 1600–1610.

Wulff, S. (2006). Go-V vs. go-and-V in English: A case of constructional synonymy? In S. T. Gries & A. Stefanowitsch (eds.), *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis*, Berlin: De Gruyter Mouton, pp. 101–126.

Yang, S.-y. (2014). L1 Transfer and Chinese as Second Language Learners' Comprehension of Noun-Noun Compounds. *US-China Foreign Language* 12(12), 953–969.

Yannakoudakis, H., T. Briscoe & B. Medlock (2011). A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT '11)*. Portland, Oregon, pp. 180–189.

Young, R. (1991). *Variation in Interlanguage Morphology*. Theoretical Studies in Second Language Acquisition. New York: Peter Lang Publishing.

Appendices

Appendix A

Analysing Subject (Pronoun) Realization: Underlying Data

In Part III, we present and evaluate a variationist approach to NLI, with Chapter 10 being dedicated to our qualitative explorations in that context. In this appendix, we provide the underlying data, used as basis for the qualitative analyses employing *subject (pronoun) realization* features, explored in Section 10.3.1. Our discussions primarily focus on patterns indicative for L1 German. In order to support analyses beyond what is included in Section 10.3.1, in this appendix we provide some information for all 11 L1s represented in our data set. The presented tables mostly contain real numbers. These numbers are logistic regression weights, assigned by the different *one-vs.-rest* classifiers corresponding to the 11 L1s. See Chapter 10, and in particular, Section 10.3.1 for further details.

variant	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR
s+”	-0.00405	-0.00306	-0.00436	-0.00448	0.01908	-0.00513	-0.00454	0.01104	-0.00556	-0.00522	-0.00450
s+cc	0.17705	-0.07437	-0.11533	-0.03461	0.04396	-0.08982	-0.07131	0.03264	-0.04780	0.09950	-0.07043
s+cd	-1.64384	-2.24032	-1.69237	0.54545	2.98573	-3.74757	-2.99539	-1.33057	-0.97783	0.12972	-0.12134
s+dt	-0.61902	-3.60778	-2.60636	1.06252	-0.63598	0.15399	-5.76571	-4.92418	-0.05210	-0.59737	-1.17003
s+ex	-3.46071	-1.40157	-5.42837	0.71296	-3.64654	-1.98968	1.47734	-0.40917	-3.43334	-2.17907	-0.95120
s+in	-1.64273	1.31372	-1.24003	-2.96388	0.16502	-1.26002	-0.88144	-1.31524	-1.92904	-0.75012	-2.77445
s+jj	0.25615	-0.22293	-0.50613	-1.67939	-0.63676	-0.09803	-0.17897	-0.52636	-0.61857	-0.00881	-0.93425
s+md	0.07532	0.29221	0.20315	-0.43732	-0.10445	-0.03042	-0.25464	-0.06327	-0.07608	0.00252	-0.28794
s+nn	-3.43583	-1.65420	-2.52616	-2.64839	-1.33169	-3.55484	-3.12926	-2.25691	-3.59049	-1.66589	-2.34013
s+pdt	-0.02002	-0.00329	-0.02376	-0.00138	0.02201	0.01719	-0.03283	-0.02572	0.07971	-0.03715	-0.00600
s+pos	0.01717	-0.00308	-0.00394	-0.00518	-0.00331	0.01554	-0.00460	-0.00362	-0.00502	-0.00505	-0.00435
s+prp$	0.16336	-0.08322	-0.07064	-0.22278	0.13797	-0.22926	-0.17719	-0.16041	0.06024	0.31366	-0.11355
s+prp	-2.12939	-2.53438	-2.08588	-2.72819	-3.97227	-1.88537	-1.33846	-2.10908	-1.98835	-3.53544	-2.70173
s+rb	0.29664	-0.27473	-0.21958	-0.22070	-0.10084	-0.22763	-0.66058	-0.41571	0.06732	0.35256	-0.21828
s+rp	0.01590	-0.01349	0.00675	0.00223	0.00334	0.00376	-0.01738	-0.01511	-0.02136	-0.02393	0.03737
s+to	-0.77108	0.15271	0.10351	-0.43876	-0.86293	0.03211	1.20315	-0.16867	-0.94031	-1.49317	0.25028
s+vb	0.81314	-1.73370	-1.01957	-3.55344	-1.41254	-4.19477	-2.49338	-0.30217	-3.39537	-1.85422	1.50640
s+wdt	1.05729	-6.17707	-3.45365	-1.35434	0.42675	1.85704	-7.80641	-7.02743	2.76451	0.24023	-2.63813
s+wp	-1.70801	-3.36306	0.21385	-1.03512	0.42223	-3.38595	-2.50008	-1.28082	-1.95515	-0.72309	-2.30098
s+wrb	0.13113	-0.03941	0.07770	-0.11923	-0.07338	0.03978	-0.15631	-0.03405	0.12822	-0.03419	-0.10865

Table A.1: Subject realization in terms of POS (ES1)

variant	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR
s+”	-0.00454	-0.00953	-0.00161	-0.00195	0.08520	-0.00831	-0.00099	-0.00511	-0.02584	-0.00282	-0.00282
s+cc	-0.01126	-0.07348	-0.05559	0.13763	-0.02713	-0.07757	0.04649	0.05224	-0.07479	0.10916	-0.07144
s+cd	0.00724	-0.43101	-0.33316	-1.02729	-0.22435	-0.95843	-0.93512	0.42103	0.43441	-0.39697	1.86405
s+dt	0.10572	-0.51373	0.00892	0.06277	-0.59888	1.09116	-0.96800	-1.03235	0.91310	-0.47445	-0.52616
s+ex	-0.87564	0.09558	-2.92349	1.11109	-0.90642	-0.53736	1.95560	0.22900	-0.71780	-1.06055	-0.15793
s+in	0.27215	0.36964	-0.68160	-0.33372	0.12863	0.64360	-0.14577	-0.54325	0.12889	-0.13478	-0.45237
s+jj	-0.53612	0.59337	-0.08418	-0.70093	-0.86201	0.41192	-0.28803	0.27154	0.06032	0.56107	-0.51620
s+md	0.20446	0.93503	0.18342	-0.47992	-0.36800	0.15576	-0.07053	-0.27440	0.14807	-0.05784	-0.39437
s+nn	-0.44336	-0.54890	0.30516	-0.05194	0.74942	-0.23929	-0.67068	-0.46313	-0.26188	0.10566	-0.81154
s+pdt	-0.02156	0.03506	-0.01179	0.02831	-0.00143	-0.01576	-0.02699	-0.01695	0.13968	-0.05654	-0.01884
s+prp$	0.34290	-0.47529	0.13695	-0.07604	0.13439	-0.19010	-0.16948	-0.29751	-0.10594	0.41692	-0.13339
s+prp	-0.32984	-0.71855	0.34246	0.52795	-0.35944	-0.44408	0.07620	-0.03388	0.04808	-0.70963	-0.38093
s+rb	-0.32831	0.10559	-0.15428	-0.07817	-0.14835	0.10267	-0.19043	-0.06771	0.33230	0.49490	-0.11526
s+rp	-0.02168	-0.00229	0.04613	-0.01371	-0.00051	-0.00490	-0.00147	-0.00130	-0.02032	-0.00020	-0.00046
s+to	-0.77463	0.61258	0.12300	-0.27929	-0.79608	-0.02623	2.18666	0.43874	-1.13134	-1.36825	0.13574
s+vb	1.92997	-0.13756	0.36586	-1.60336	-0.98008	-0.27456	-0.15095	0.45641	-1.57884	-1.67066	0.70099
s+wdt	0.14192	-2.03412	-0.72453	-0.37203	0.61270	0.02175	-1.45801	-0.94444	-0.46966	1.91874	-0.41715
s+wp	-0.40494	-0.45163	0.55773	-0.60226	1.31146	-1.19687	-0.96771	0.21537	-0.78966	0.86361	-0.46024
s+wrb	-0.00665	-0.07084	0.25772	0.20556	-0.26888	0.18906	-0.28088	-0.20394	0.32550	0.07324	0.10790

Table A.2: Subject realization in terms of POS, verb lemma *be* (ES2)

variant	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR
s+dt	0.41684	-0.06549	-0.25088	-0.11680	-0.23022	-0.06349	-0.00938	-0.05157	0.32002	-0.04192	-0.12742
s+ex	-0.70538	0.07702	-0.61784	0.68052	-0.35526	-0.81127	-0.64226	0.72215	0.56763	-0.86784	1.28835
s+in	0.63504	-0.06303	0.48949	-0.15522	-0.06450	-0.07898	-0.02402	-0.03481	-0.26281	-0.08045	-0.30233
s+jj	-0.02861	-0.00116	0.36547	-0.00082	-0.01489	-0.01206	-0.00056	-0.00587	-0.12668	-0.00255	-0.00434
s+nn	-0.62718	0.14207	0.52159	0.06026	0.04856	0.30278	-0.20074	-0.01015	-0.06879	0.64254	-1.51801
s+prp	-0.09750	-0.47774	0.99739	-0.25587	-0.11122	-0.03952	0.24912	-0.16878	-0.70147	-0.11107	0.41529
s+vb	-0.00622	-0.05628	-0.12542	-0.01056	-0.00721	-0.01999	-0.06650	-0.02197	0.59300	-0.02405	-0.01204
s+wdt	0.17484	-0.43686	-0.64098	0.15672	0.88655	-0.27191	0.01922	0.23192	-0.04948	-0.03748	-0.21961
s+wp	-0.15450	-0.04467	0.11230	-0.15359	-0.30278	0.73662	-0.02955	-0.32559	-0.18349	-0.07218	0.42017

Table A.3: Subject realization in terms of POS, verb lemma *exist* (ES2)

variant	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR
s+in	0.34822	1.03610	-0.74135	0.02066	0.18730	-0.24620	-0.74268	-0.17179	-0.44713	-0.46237	0.04060
s+nn	-0.24548	-0.41753	-0.38959	-1.17540	-0.00525	0.24317	-0.73796	0.45959	0.71492	0.49264	0.74304
s+prp	0.33000	-0.76819	0.90520	-0.36099	-0.88717	0.15121	0.80562	-0.36262	-0.36004	-0.66981	0.15709
s+wp	-0.20191	-0.26373	-0.60686	0.73016	0.90598	-0.35610	0.56595	-0.47039	-0.22084	0.08128	-0.09418

Table A.4: Subject realization in terms of POS, verb lemma group G1={*accept, define, govern, respect*} (ES3)

variant	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR
s+nn	-0.71641	0.08709	-0.32992	-0.62367	0.87681	-0.62042	0.95862	-0.29392	0.08785	-1.16697	0.77474
s+prp	-0.16140	0.11350	-0.08596	-0.50491	-0.09357	0.55225	-0.32691	-0.52813	0.58926	0.37618	-0.24972
s+wdt	0.95619	-0.48942	-0.27130	0.84539	0.16100	-0.45597	-0.24958	-0.14628	-0.73027	-0.12796	-0.14366
s+wp	-0.24259	-0.07072	-0.45657	0.65411	0.67479	0.30089	-0.19376	-0.03730	-0.28035	-0.10374	-0.33901

Table A.5: Subject realization in terms of POS, verb lemma group G2={*differentiate, fly, relate, suit*} (ES3)

variant	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR
s+nn	0.06068	0.35871	-0.19061	0.10974	-0.06837	-0.41548	-0.30458	0.16171	-0.23161	-0.47842	0.22808
s+prp	0.67361	-0.22364	-0.26278	-0.22478	-0.32354	0.10643	-0.22067	-0.05843	0.11820	0.78706	-0.30472
s+wdt	-0.52089	-0.42616	0.45628	1.17197	0.69487	0.65176	-0.48643	-0.60506	-0.56784	-0.43760	-0.75018

Table A.6: Subject realization in terms of POS, verb lemma group G3={*arise, emerge, stem*} (ES3)

variant	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR
s+nn	-0.72589	0.66676	-0.72346	0.02072	-0.31465	-0.37493	0.19478	0.43070	0.23254	-0.58889	0.21131
s+prp	-0.69091	-0.90783	0.86781	0.12122	-0.10858	-0.51724	-0.17726	0.11124	0.34003	-0.26544	-1.39723

Table A.7: Subject realization in terms of POS, verb lemma group G4={*admire, avoid, fear, worry*} (ES3a)

group id	# lemmas	lemma list
G1	166	adopt, advertise, affect, affirm, aim, analyze, announce, appeare, approach, arent, argue, aske, assign, avoid, base, bear, bee, blame, bring, care, change, claim, close, communicate, complain, complete, confirm, consist, constitute, contain, continue, convey, convince, costs, count, cover, critize, cut, damage, decrease, decribe, deliver, determine, develop, die, differ, diminish, disappear, distinguish, drop, e, emphasize, enable, encourage, end, enjoy, ensure, envolve, equal, establish, estimate, evaluate, exagerate, exaggerate, exclude, exist, expand, explain, explaine, express, fact, fear, feature, fit, focuss, force, form, fulfill, great, hase, haven, hesitate, heve, hold, illustrate, increase, indicate, infer, influence, insist, involve, is, judge, knowledge, lack, last, lay, lead, let, lie, link, loose, make, mark, matter, mature, mislead, modify, nurture, occur, offer, perform, pick, place, please, point, pollute, present, pretend, promote, proove, prove, provide, push, raise, react, reflect, regret, regulate, rely, remain, represent, require, reserve, respond, retain, retire, reveal, reward, rise, rule, run, schedule, set, settle, show, sit, solve, sould, sound, specilize, spendt, step, stimulate, store, stress, suffer, summarize, taugh, teach, transfer, trie, triumph, vary, whre, works
G2	100	ahve, allow, appeal, appear, apply, aspect, assert, attract, be, become, becuase, begin, broaden, calculate, cast, catch, cause, characterize, charge, chose, come, compete, confine, connect, contribute, control, create, demonstrate, demostrate, derive, deserve, develope, educate, engage, enlarge, evoke, evolve, examine, experienced, fight, focus, give, grow, hang, happen, help, hire, hurt, implement, imply, include, induce, inform, invade, invest, justify, list, mantain, may, might, operate, paint, participate, permit, play, pose, predict, produce, question, red, reduce, regard, request, respect, satisfy, say, seem, shold, shoul, should, start, state, stick, sustain, tell, tend, thank, throw, try, understnad, undertand, utilize, view, wether, wich, will, willl, work, woul, young

G3	312	've, —-, absorb, accomplish, achieve, achive, acknowledge, acquire, acumulate, add, adjust, admire, agree, agrre, al, allway, am, annoy, answer, appreciate, approve, arrange, arrive, arround, ask, aspire, assimilate, associate, assume, assure, attend, aware, bare, bege, beleave, belief, believe, belive, belong, bet, book, build, buy, ca, call, can, cant, cannot, cant, carry, check, cherish, choose, clarify, clean, climb, coast, combine, commit, compare, consider, construct, cook, cost, coul, could, cross, d, dare, debate, decide, define, delay, depen, depend, describe, design, desire, destroy, didnt, disagree, discover, discuss, dislike, dive, do, doe, doesn, don, don-t, dont, dream, drive, earn, earnd, eat, effect, emerge, employ, encouter, entail, enyoy, expect, expend, experience, explore, expose, extend, face, fail, fall, fancy, feed, feel, felt, fetch, find, finish, fly, focuse, follow, forget, found, fun, gain, generalize, get, go, graduate, guess, guide, gurantee, habe, hav, have, hear, hide, hope, ie, ignore, imagine, impact, improve, infact, intend, interact, introduce, invent, invert, invite, ist, jump, keep, kill, knock, know, konw, learn, learnd, learnt, leave, life, like, limit, listen, live, look, lose, love, maintain, mean, measure, meet, memorize, mention, mermorize, migth, mind, miss, move, mus, must, narrow, need, note, notice, observe, omit, open, order, organize, ought, own, pay, perfom, plan, ponder, possess, practic, prefer, prefere, preffer, prepare, preview, product, promove, propose, purchase, put, quit, range, reach, read, realise, realize, reallz, recall, receive, recive, recognize, record, refere, refresh, refuse, reject, relate, relax, release, remeber, remember, repeat, resolve, result, return, review, risk, s, save, saw, scare, see, seek, sell, selves, share, shoot, shoud, sill, sing, skill, slow, smell, smile, some, spanish, speak, specialize, spend, stand, star, starte, stay, stop, study, succe, succeed, success, support, suppose, suspect, switch, t, take, talk, taste, thing, think, tink, travel, travell, travelle, travelling, treat, trust, turn, understand, undestand, urge, use, usuallz, value, visit, wait, wake, walk, want, wash, watch, wear, wil, win, wish, witness, wo, wolud, wonder, wont, worth, would, wound, write, wrtite

G4	65	act, ate, behave, benefit, boast, break, cite, command, concentrate, concerned, conclude, conduct, consume, convice, copy, creat, deny, devote, display, divide, draw, enhance, equip, exploit, fell, generate, grasp, human, interest, join, lanch, laugh, manipulate, match, meke, ment, oblige, obtain, occupy, occure, oppose, opt, party, pass, persist, prevent, proceed, recieve, refer, remind, replace, search, seat, shoule, showe, spread, stays, stem, subject, suggest, summerize, tempt, touch, underline, waste
G5	68	accept, accumulate, advocate, ae, aloud, assist, bad, bougth, burn, challenge, collect, complement, concern, concur, cope, corrispond, deal, dedicate, demand, drink, encounter, enter, entertain, experiment, favour, feeling, figure, finance, hace, handle, happend, hate, idea, innovate, ll, manage, marry, master, motivate, organise, outline, outweight, perceive, perfer, persuade, polluate, profit, promise, pursue, recommend, reproduce, requiere, send, serve, shall, shouldn, sign, sleep, split, struggle, survive, target, tast, test, undergoe, undertake, worry, wouldn

Table A.8: Verb lemma groups yielded by the grouping technique at the best cut-off, employing *s+prp* and *s+nn* as features

POS	lemma
dt	a, ahe, all, an, andi, another, anotherone, any, both, e, each, either, ell, enjoy, every, few, i, less, many, more, most, neither, niether, no, one, onother, several, some, such, th, that-is-to, that, thay, the, theer, ther, thery, these, thid, thier, this, those, thre, thy, u, wich, yo
ex	althoughthere, everyday, i, tere, thanthere, thatthere, theere, ther, thera, there, therea, thereare, thereb, therefor, therer, theres, theri, thingsd, threr, threre, tthere, u, wich, wish, youth
prp$	her, his, its, iy, my, our, their, thier, your
prp	't, ae, aim, ar, do, don, e, everone, ey, f, ge, h, haw, he, her, herself, hi, him, himself, ho, htey, hu, human, hw, i, iam, ido, ie, ifself, ii, im, innerself, it, itsself, itself, l, ll, lt, me, medium, morthey, myself, oe, one-self, one, oneself, ot, ou, ouer, ours, ourselves, s, self, she, si, som, sur, t, te, tha, thay, theire, theirl, theirs, theirselves, them, themself, themselfes, themselves, ther, thet, thetyou, they, theycan, theyh, theyr, theyre, theyself, thez, thezbuy, thim, thr, thre, threre, thy, ti, tou, tu, tyey, u, us, ve, vrey, w, we, wer, y, yiu, yo, yon, yopu, you, youn, young, youngster, yourself, yous, yoy, ypu, yself, yu, yuo, ze, zou
wp	eho, rhat, tho, waht, wgo, whant, what, whit, who, whoever, whom, whome, whon, whorever, zho

Table A.9: Subject pronoun realization in terms of POS and the corresponding lemmas

variant	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR
s+dt+both	-0.36849	-0.00826	-0.00043	0.12267	0.32852	-0.42920	-0.64558	-0.66288	0.10737	-0.34499	-0.21987
s+dt+neither	-0.11974	-0.09021	-0.12043	0.39312	-0.02283	-0.00514	0.12972	-0.13418	-0.00920	-0.13849	-0.09358
s+dt+the	-0.01755	-0.13258	-0.22144	0.18003	-0.09257	-0.06011	-0.18820	-0.11752	0.18354	0.01908	-0.04723
s+dt+this	-2.62970	-6.25349	-4.10531	0.66282	1.21724	0.22669	-5.88583	-6.21648	-0.98196	0.24463	-1.66916
s+ex+threre	-0.03605	-0.02655	-0.02435	0.18557	-0.03743	-0.02888	-0.02923	-0.02614	0.03456	-0.05625	-0.03911
s+prp+one	-1.26545	-0.84933	-0.60441	1.00502	1.52349	-2.23470	-1.85610	-1.33097	-0.50363	-0.42346	-0.64523
s+prp+zou	-0.10141	-0.06817	-0.05503	0.48866	-0.06246	-0.11070	-0.08477	-0.07295	-0.10192	-0.04964	-0.07909

Table A.10: Subject pronoun realization in terms of POS and lemma, most indicative variants for L1 German (ES1)

variant	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR
s+dt+neither	-0.09268	-0.10363	-0.32101	0.51783	-0.08281	0.25861	0.18095	-0.19635	-0.06481	-0.03214	-0.14078
s+dt+this	-0.51976	-1.01653	-0.81452	0.76891	-0.24343	1.85210	-0.06362	-1.07621	0.01571	0.18859	-0.19854
s+dt+those	-0.07983	-0.28552	-1.03367	0.32731	-0.65663	-0.44602	0.30241	1.13853	0.13196	0.72143	-0.20945
s+ex+there	-0.42454	0.20387	-1.48443	0.63548	-0.44519	0.00509	0.97456	0.16865	-0.41611	-0.46944	0.06684
s+ex+therefor	-0.00984	-0.00101	-0.00779	0.10281	-0.00079	-0.00104	-0.00011	-0.00063	-0.02061	-0.02170	-0.00413
s+ex+threre	-0.12444	-0.03765	-0.06699	0.50243	-0.04651	-0.07483	-0.11939	-0.12594	0.19594	-0.11995	-0.10903
s+prp+me	-0.07834	0.00380	-0.11430	0.11247	-0.04750	0.09963	-0.10361	-0.09928	0.45165	0.13652	-0.07644
s+prp+she	-0.39546	-0.30209	-0.14554	1.25217	-0.04547	-0.32976	0.48962	0.48225	-0.19369	-0.35184	0.66853
s+prp+they	-0.17502	0.03418	0.01440	0.15116	-0.29099	-0.35241	0.12406	0.02998	-0.05787	-0.11352	-0.24394
s+prp+thy	0.16123	-0.06503	-0.02184	0.16688	-0.04594	-0.00591	-0.01585	-0.07878	-0.05147	-0.06308	-0.06118

Table A.11: Subject pronoun realization in terms of POS and lemma, most indicative variants for L1 German, verb lemma *be* (ES2)

variant	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR
s+ex+there	-0.66546	-0.11986	-1.17934	0.65026	-0.23767	-1.09232	-0.81186	0.47057	0.42265	-0.59560	0.82588
s+prp+they	-0.03878	-0.37427	0.24255	0.46435	0.18650	0.02975	-0.23905	-0.05707	-0.02852	-0.47916	0.35419

Table A.12: Subject pronoun realization in terms of POS and lemma, most indicative variants for L1 German, verb lemma *exist* (ES2)

variant	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR
can	-0.61138	-0.50358	0.27254	0.88296	0.74032	-1.68125	-1.39493	-0.63307	0.19848	0.14050	-0.16176
could	-0.09275	-0.08890	-0.25088	0.92654	-0.21409	0.24943	-0.04047	-0.10479	-0.20303	-0.22847	-0.17044
have	0.46568	-0.34142	-0.14146	0.83522	0.46567	-0.16349	0.04796	-0.18328	-0.11872	-0.49747	-0.14691
might	-0.46400	-0.64447	0.26713	0.60643	0.05729	0.08457	-0.60535	-0.31490	0.37872	-0.62012	-0.07818
must	-0.73054	-0.36367	0.13387	1.11058	0.57812	-0.32268	-0.50335	-0.28049	-0.19532	-0.14634	-0.27844
should	-0.35981	-0.35597	0.16620	0.91388	0.78337	-0.31315	-0.36417	-0.23711	-0.22904	-0.46241	-0.21202
sould	-0.05185	-0.00566	-0.04061	-0.05251	-0.01635	-0.05642	-0.02373	-0.03105	-0.14485	-0.01517	0.35518
will	0.16319	-0.08112	-0.06085	-0.07115	-0.10704	-0.02408	-0.04993	-0.05597	0.76250	-0.11033	-0.24232
would	-0.55876	-0.37842	0.70985	-0.23196	-0.45328	-0.23894	-0.36997	-0.34434	0.49162	0.51244	0.41900

Table A.13: All verb lemmas realized with the variant *s+prp+one* (ES2)

variant	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR
s+dt+this	-0.30049	-0.99229	-0.83545	0.85315	0.63560	0.13287	0.27584	-0.23888	0.52949	-0.23121	-0.26593
s+prp+it	0.19594	-0.33632	0.88408	-0.77457	-1.15322	0.05794	-0.70213	0.42376	0.31655	-0.87984	-0.47943
s+prp+they	0.15187	0.13404	0.80298	-0.18919	-0.63037	0.11609	-0.46860	0.06583	-0.06046	0.03789	-0.07361

Table A.14: Subject pronoun realization in terms of POS and lemma, verb lemma group G1={*allow, imply, motivate, prevent*} (ES3)

variant	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR
s+dt+this	0.13971	-0.68110	-0.24240	0.18056	-0.54974	-0.46748	-0.26081	-0.30250	-0.34763	0.33115	-0.11665
s+prp+he	-0.06660	-0.04732	-0.16380	-0.26268	-0.05590	0.25135	-0.02115	-0.14498	-0.03189	-0.00832	-0.14071
s+prp+i	-0.11317	-0.10277	0.75029	0.31959	-0.04989	-0.30255	-0.13579	-0.08269	-0.11844	-0.13671	0.23746
s+prp+it	0.16098	-0.12736	-0.64687	-0.43717	1.01376	-0.63223	-0.49126	0.10077	-0.52583	0.51186	0.24634
s+prp+they	-0.41704	-0.08105	-0.26284	0.48175	0.76820	-0.03388	-0.29500	0.04838	-0.11175	0.11637	-0.45286
s+prp+we	0.23771	0.82374	-0.20685	-0.25688	-0.35380	0.15579	0.30306	-0.17731	-0.12311	-0.17136	-0.31391
s+prp+you	-0.35120	-0.25964	-0.45233	0.11477	-0.12110	0.40500	-0.12322	-0.10837	0.38229	-0.10196	0.67048

Table A.15: Subject pronoun realization in terms of POS and lemma, verb lemma group G2={*count, increase*} (ES3)

variant	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR
s+dt+all	-0.15041	-0.20505	-0.08959	-0.01537	0.55206	0.77885	-0.04808	-0.14299	-0.06856	-0.03877	-0.28129
s+dt+many	0.68432	-0.14279	-0.13347	-0.04367	-0.10756	0.60852	-0.06603	-0.12122	-0.14094	-0.13917	-0.03673
s+dt+most	-0.23808	0.63716	-0.19839	-0.25388	-0.19232	-0.21849	-0.11636	0.40079	-0.33497	-0.04957	0.66626
s+dt+that	1.06366	-0.62829	-0.28967	0.24304	-0.47813	-0.12873	-0.30329	0.36182	-0.65534	0.09570	-0.88323
s+dt+these	-0.46605	-0.18733	-0.16548	-0.17484	0.08421	-0.37439	-0.17755	0.81900	0.44136	0.18857	-0.39298
s+dt+this	-0.24283	-0.87913	-0.58713	-0.33777	0.07348	0.27916	-0.12552	-0.70520	0.52647	0.52163	0.28128
s+ex+there	-0.14189	-0.18631	-0.31012	0.60731	-0.22771	-0.14579	-0.32076	0.27194	-0.22898	0.79055	-0.37727
s+prp+i	0.27113	0.99169	-0.50489	0.05838	0.17427	-0.46381	0.11039	-0.88570	-0.92004	1.00634	0.27023
s+prp+it	-0.36532	-0.10225	0.11002	-0.05980	0.23113	0.53918	0.26188	-0.14654	-0.41221	0.22286	-0.70947
s+prp+they	-0.29669	-0.29388	-0.02259	0.01842	-0.37315	0.51493	0.22457	-0.43582	0.26580	-0.30033	0.27031
s+prp+we	-0.14117	-0.70122	0.59473	-1.02359	-0.34589	-0.54082	0.50002	0.34694	-0.36377	0.26552	0.48477
s+prp+you	0.75705	-0.11147	0.55362	-0.46814	-0.27782	0.28870	-0.34096	-0.79393	0.14198	0.06464	-0.06621
s+wp+what	-0.11383	0.19162	0.80288	0.05008	0.01217	0.19655	-0.34591	-0.58933	-0.29366	-0.42561	0.70398
s+wp+who	-0.38081	0.02718	-0.34935	0.65119	0.72961	-0.30000	-0.91926	-0.15566	0.36388	-0.00583	-0.15256

Table A.16: Subject pronoun realization in terms of POS and lemma, verb lemma group G3=$\{$*give, happen*$\}$ (ES3)

Appendix B

Analysing Nominal Modification: Underlying Data

In Part III, we present and evaluate a variationist approach to NLI, with Chapter 10 being dedicated to our qualitative explorations in that context. In this appendix, we provide the underlying data, used as basis for the qualitative analyses utilizing *nominal modification* features, explored in Section 10.3.2. Our discussions primarily focus on patterns indicative for L1 German. In order to support analyses beyond what is included in Section 10.3.2, in this appendix we show some information for all 11 L1s represented in our data set. The provided tables contain real numbers. These numbers are logistic regression weights, assigned by the different *one-vs.-rest* classifiers corresponding to the 11 L1s. See Chapter 10, and in particular, Section 10.3.2 for further details.

variant	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR
$+N	-0.00531	-0.00652	0.00524	-0.00568	0.02149	-0.00633	-0.00734	0.00874	0.00618	-0.00804	-0.00689
”+N	-0.01041	0.00873	-0.02310	-0.01146	0.03794	-0.01196	-0.00040	-0.01020	-0.00401	-0.02367	0.03323
cc+N	0.44644	-0.32515	-0.09959	-0.04882	0.32817	-0.18988	-0.27885	-0.26011	-0.14904	0.00871	-0.22320
cd+N	-0.21239	-0.63801	-0.80216	-1.90631	-4.63770	-2.25503	0.66560	0.90195	-2.28120	-3.02184	-2.45566
ex+N	1.78157	-0.86350	-0.54497	-0.10467	0.38238	-0.78728	-0.76408	-0.77877	-0.87778	1.06159	-0.76100
N+”	0.01021	-0.01730	-0.00265	-0.01559	-0.00326	0.05201	-0.01934	-0.02045	-0.01567	-0.00466	0.01548
N+cc	-0.70882	-2.31857	-4.50652	-2.75026	-2.07073	-4.61564	-0.95353	-1.16412	-1.92616	-2.45991	-3.25486
N+cd	-0.09218	-0.07894	0.02208	-0.12014	-0.09723	0.06901	-0.01667	-0.07902	-0.02332	0.06637	-0.05393
N+ex	0.04606	-0.03180	-0.00290	-0.00820	-0.00871	-0.01524	-0.01144	-0.03157	-0.00088	0.05919	-0.03495
N+Nyph	-0.29962	-0.54742	-0.14451	1.82999	0.01058	-0.20597	-0.66834	-0.34944	-0.24975	-0.36389	-0.08937
N+in	-4.49346	-4.67401	-1.66808	-2.31662	-0.12790	-1.59868	-5.36667	-6.03452	-2.54325	-2.04924	-1.56280
N+jj	-0.20458	-0.71374	-0.49209	-0.20395	0.11138	0.84728	-1.32067	-1.77237	-0.39511	-0.93490	-0.97908
N+jjr	-0.17006	-0.05950	0.05472	-0.03255	-0.24830	0.09808	-0.09309	-0.06607	0.16444	-0.21050	0.10311
N+jjs	-0.00769	-0.04869	0.00642	-0.01531	-0.00477	0.01471	-0.04204	-0.01629	0.07254	-0.03320	0.01474
N+ls	-0.00501	-0.00482	-0.00389	-0.00439	0.01091	0.00617	-0.00584	-0.00554	0.00513	-0.00542	0.01508
N+md	3.73144	-2.06520	-0.42904	-0.83146	-1.06126	-0.53484	-5.82143	-5.50606	1.47014	0.90559	-1.95257
N+nil	0.07987	-0.09833	-0.07138	-0.03677	0.00782	-0.02522	-0.13390	-0.08131	-0.03707	0.23097	0.00945
N+nn	-0.43715	-1.02600	-0.13794	-2.28323	-0.64855	0.85037	-1.03414	-1.86814	1.31170	-2.15378	-3.84231
N+nnp	-0.49690	-0.24540	-0.26482	-0.59980	0.05361	-0.35170	-0.21851	0.17724	0.01103	-0.27741	-0.40856
N+nnps	-0.01582	-0.00905	0.03060	0.00401	-0.00119	-0.01668	0.04560	-0.02607	-0.00552	-0.01853	-0.01770

N+nns	-0.73471	-1.59051	-0.47564	-1.09475	-0.54548	-0.00770	-1.78057	-1.94305	1.80629	-0.40982	-0.09674
N+pdt	-0.00455	-0.00379	-0.00404	0.02302	-0.00397	-0.00431	-0.00371	-0.00355	-0.00523	-0.00433	0.01169
N+pos	-2.44814	0.77991	-0.81408	-2.05366	-1.23938	0.02074	-0.01311	2.95910	-1.35290	-2.53899	0.41299
N+prf	-0.00073	-0.00101	-0.00067	-0.00069	-0.00093	-0.00060	-0.00117	0.00886	-0.00070	-0.00114	-0.00096
N+prp$	0.01007	0.01944	0.00397	-0.01796	-0.00424	-0.01900	0.01448	-0.02288	0.01429	0.00565	-0.02099
N+prp	0.26640	-0.32641	0.04044	-0.15288	-0.18231	-0.04431	0.00375	-0.40566	-0.15072	-0.07232	-0.44399
N+rb	-0.36769	-0.62086	-0.79540	0.44555	-0.02180	-0.96602	-1.27191	-0.75095	-0.11919	0.01364	-0.46408
N+rbr	-0.12529	0.08054	-0.06570	0.00592	-0.11434	-0.14021	0.03879	0.17576	-0.12170	-0.31727	-0.16497
N+rbs	-0.01892	-0.03053	0.00545	0.05812	0.02094	-0.03059	-0.03334	-0.03234	-0.00894	-0.00326	0.01904
N+rp	-0.06258	-0.00057	-0.06282	0.01612	0.03399	-0.00089	-0.02176	-0.05962	0.00287	0.04721	-0.05462
N+sym	-0.00491	0.00982	-0.00431	0.00174	-0.00437	-0.00468	-0.00463	-0.00409	-0.00622	0.01903	-0.00437
N+to	-0.71485	0.44075	0.32380	-4.04311	-4.48257	-2.18513	-1.57940	-1.75778	0.52737	-4.61207	-2.70673
N+uh	-0.02301	0.04060	0.03342	-0.02033	0.02073	-0.01881	-0.02554	-0.00272	0.00042	-0.02204	-0.02045
N+vb	-0.02766	0.00262	0.09433	-0.13222	-0.05381	0.04392	-0.23623	-0.03842	0.16651	-0.14906	-0.21359
N+vbd	0.93355	-2.45851	-1.50264	-1.63049	-1.49617	-0.34642	-0.24122	-0.93513	0.55349	-0.33494	-1.99511
N+vbg	-0.46481	-1.31781	-1.17397	-0.43312	1.55869	-1.37399	-2.68548	-1.42442	-0.64508	1.86979	-0.46081
N+vbn	-1.32573	0.41111	-1.03189	-1.88197	0.20611	-0.22753	-1.62219	-0.75121	-0.19644	0.31940	-2.09611
N+vbp	0.13088	-2.71014	-1.94583	-1.23251	-5.80083	-0.33884	-2.41653	-4.86991	3.76251	-3.37776	-3.67671
N+vbz	0.03932	-5.76092	-1.08003	-0.55957	0.32921	-0.60919	-4.00166	-3.93348	1.17416	0.64137	-1.13793
N+wdt	0.04015	-0.08965	-0.06657	-0.07990	-0.01150	0.07049	-0.00930	0.04835	0.03202	-0.02530	-0.05161
N+wp$	0.00986	0.00836	0.01709	-0.00433	-0.00447	-0.00471	-0.00498	-0.00486	-0.00429	-0.00482	-0.00514

N+wp	0.06648	0.00816	0.02411	-0.02171	-0.01503	-0.04342	0.00013	-0.03023	-0.02156	-0.01417	-0.01706
N+wrb	0.03663	0.00893	-0.03679	-0.01567	-0.04685	-0.01631	-0.01056	0.00055	0.05293	-0.03532	-0.05901
hyph+N	-0.02106	-0.01121	0.07096	0.00769	-0.00388	-0.03594	-0.05738	-0.03262	0.06202	-0.01939	-0.05795
in+N	-0.20924	-0.22717	-0.77594	-0.45523	-0.67160	-0.50492	-0.20469	0.40279	-0.02490	-0.35156	-0.24479
jj+N	-3.85222	-1.68635	-2.74330	-1.80175	-4.86849	-2.00516	-1.86635	-1.91191	-4.07778	-3.32165	-2.97768
jjr+N	-1.30451	-1.98027	-3.35575	0.01308	0.00992	-5.22616	0.18842	-0.76166	-2.60874	1.17242	-2.68180
jjs+N	-0.08652	-0.59154	0.02127	-0.87845	-1.95638	-0.27795	-0.40518	-0.91380	1.21250	-2.60566	-0.67785
ls+N	-0.03042	-0.02756	-0.02539	0.11443	0.00941	0.00409	-0.03695	-0.05767	0.01279	0.02296	-0.05520
md+N	-0.01460	0.03482	-0.00874	-0.01149	-0.00942	-0.00964	0.01609	-0.02844	-0.00775	-0.01058	0.01260
nil+N	0.03058	-0.03420	-0.03593	-0.01354	0.00781	-0.00394	-0.06784	-0.03372	-0.00389	0.05936	-0.00009
nn+N	-1.93216	-1.14763	-5.59898	-6.66468	-1.59213	-6.94177	0.55528	0.83905	-4.94284	-1.47664	-2.48040
nnp+N	-2.39740	-2.17440	-1.86507	-2.85474	-0.14163	-2.77729	-1.68899	-1.31878	-1.49088	-0.30329	-1.92951
nnps+N	-0.07693	-0.09834	0.00969	0.00866	-0.07157	-0.01008	0.05364	-0.08036	0.07333	0.01539	0.04447
nns+N	0.10207	-0.89722	0.17140	-1.93122	-1.84066	-0.87411	-1.93389	-1.52958	1.17267	0.26649	-0.74883
pdt+N	-0.58494	-1.03194	0.42587	-0.81261	0.32919	1.34203	-3.88124	-3.16251	1.19769	1.02873	-1.85039
pos+N	0.06000	-0.05107	-0.03562	0.00106	-0.01454	0.08773	-0.04667	-0.05193	-0.02371	-0.04148	0.00683
prf+N	-0.01882	-0.04006	0.03188	-0.01923	0.05716	-0.01587	-0.04283	0.00248	0.05842	-0.03123	-0.02852
prp$+N	-0.94386	-2.50499	-2.28932	-2.96890	-2.56277	-2.91335	-2.41209	-0.78162	-3.56223	-3.86615	-2.21475
prp+N	0.46362	0.04211	-0.03916	-0.21539	-0.13433	-0.33134	-0.28202	-0.20000	0.05737	-0.18620	-0.60520
rb+N	-2.83003	-1.48804	-2.26414	0.53080	-1.94956	0.20640	-0.68504	-0.68167	-0.71363	-1.92839	-1.07011
rbr+N	-0.42951	1.19662	0.18275	-0.43640	-0.23173	-0.57726	-0.44383	-0.10194	-0.13294	-0.61021	-0.53154

rbs+N	-0.21505	-0.11446	-0.49245	-0.47829	-0.70713	-0.54194	0.56858	-0.21073	-0.65920	-0.63140	-0.55841
rp+N	0.00618	-0.00105	-0.01470	-0.01634	-0.01562	-0.01753	0.02952	-0.01289	-0.00393	0.04123	0.00809
sym+N	0.01273	-0.02992	-0.02129	-0.00916	0.05614	-0.02389	-0.01168	-0.01941	-0.02700	0.02587	0.00743
to+N	0.01126	-0.02442	-0.02867	0.00175	0.00579	0.00912	-0.03763	0.02721	-0.03092	0.02776	0.01509
uh+N	-0.01635	0.09567	-0.05399	-0.05330	0.00159	-0.05591	-0.03379	-0.04453	-0.06617	0.00626	-0.02536
vb+N	0.03599	0.11663	0.02347	-0.02467	-0.07741	-0.07181	-0.10073	0.08586	-0.05398	-0.02731	-0.14772
vbd+N	0.08354	-0.02837	-0.03978	0.04883	-0.02510	-0.08692	-0.09264	-0.02355	-0.02432	0.00385	0.03194
vbg+N	-1.68626	0.44982	-1.06180	0.89937	0.17508	-2.67423	-1.02496	0.30691	-2.65481	0.54637	0.72810
vbn+N	-1.55682	-1.99594	-0.91762	1.91154	2.25602	-1.41003	-2.15424	-0.54229	-1.28597	0.94966	-0.62455
vbp+N	-0.01784	-0.02333	-0.00964	-0.06667	-0.01627	-0.01275	-0.02897	-0.01234	0.00085	0.00916	-0.01764
vbz+N	0.04241	0.03302	-0.02249	-0.02662	-0.05373	0.01074	-0.06964	-0.01339	0.00799	0.02713	-0.02838
wdt+N	-0.19627	-0.00419	-0.03609	-0.21458	-0.24254	-0.24966	-0.01973	-0.07263	-0.26631	-0.15180	-0.05112
wp$+N	-0.11899	-0.05030	0.00981	-0.12627	0.12099	0.00368	0.02422	-0.09010	0.10159	-0.08254	0.00453
wp+N	-0.11307	0.20393	-0.21119	0.13427	-0.56442	-0.15150	0.25015	-0.03946	0.19994	-0.76527	-0.57474
wrb+N	0.10806	0.16492	-0.17419	0.14020	-0.36337	0.02519	-0.06065	-0.11630	0.13825	-0.48468	-0.38494

Table B.1: Nominal modification in terms of POS, N = head noun
(ES1)

variant	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR
N+hyph	-0.29962	-0.54742	-0.14451	1.82999	0.01058	-0.20597	-0.66834	-0.34944	-0.24975	-0.36389	-0.08937
N+rb	-0.36769	-0.62086	-0.79540	0.44555	-0.02180	-0.96602	-1.27191	-0.75095	-0.11919	0.01364	-0.46408
ls+N	-0.03042	-0.02756	-0.02539	0.11443	0.00941	0.00409	-0.03695	-0.05767	0.01279	0.02296	-0.05520
rb+N	-2.83003	-1.48804	-2.26414	0.53080	-1.94956	0.20640	-0.68504	-0.68167	-0.71363	-1.92839	-1.07011
vbn+N	-1.55682	-1.99594	-0.91762	1.91154	2.25602	-1.41003	-2.15424	-0.54229	-1.28597	0.94966	-0.62455
wrb+N	0.10806	0.16492	-0.17419	0.14020	-0.36337	0.02519	-0.06065	-0.11630	0.13825	-0.48468	-0.38494

Table B.2: Nominal modification in terms of POS, most indicative variants for L1 German, N = head noun (ES1)