A Bayesian multinomial model of near-synonymy

A corpus-based analysis of letting constructions in English

Natalia Levshina
F.R.S – FNRS, IL&C,
Université catholique de Louvain
Louvain-la-Neuve, Belgium

natalevs@gmail.com

***Abstract*—This paper is a quantitative multifactorial study of near-synonymous constructions *let* + V, *allow* + *to* V and *permit* + *to* V based on the British National Corpus. We fit a Bayesian multinomial mixed model with twenty formal, semantic, social, collostructional and other variables as fixed effects and the infinitives that fill in the second verb slot as random effects. The model reveals a remarkable alignment of variables that indicate the formal distance between the predicates, conceptual distance between the events they represent and between the speaker and the main arguments, the social and communicative distance between the interlocutors, as well as the looseness of the relationship between the constructions and second verb slot fillers. These results raise fundamental theoretical questions about the relationships between linguistic form, function and use.**

***Keywords*—Bayesian multinomial regression; letting; iconicity; near synonymy; multifactorial models**

I. Aims of the study

There has been an increasing number of quantitative studies that determine the factors that help predict the speaker's choice between functionally similar constructions, such as English phrasal constructions with varying particle placement [1], German middle field alternation [2], Finnish verbs of thinking [3], Russian verbs of trying [4], and Shanghainese topic markers [5]. However, even though these models are descriptively adequate, they often focus on the effects of the semantic, pragmatic, collocational, social and other factors separately, without considering the underlying relationships between them. In this study, we want to show that a multifactorial analysis of near-synonyms can raise fundamental questions that are relevant for general linguistics.

The object of this study is variation of three English constructions of letting: *let* + V, *allow* + *to* V and *permit* + *to* V. Examples from the British National Corpus are provided in (1):

(1) a. *I am content to let you form your own judgment of my character.* (H84)

b. *Representation 1 allows us to depict any set of pairs of coordinates.* (FNR)

c. *In this form the censor permitted the book to pass.* (B7K)

This paper investigates several dimensions of variation of the constructions: formal, semantic, cognitive, social and collostructional [6]. To the best of our knowledge, this is a first study that investigates how all these dimensions are aligned in near-synonymous constructions and which offers a discussion of this alignment from a theoretical point of view.

From a methodological perspective, this paper is innovative, as well: the effect of the twenty variables that represent the dimensions is tested with the help of a cutting-edge statistical technique, Bayesian multinomial mixed-effect regression, which is implemented in the R package *MCMCglmm* [7]. Since near-synonymy in lexicon and grammar is not restricted to pairs of functionally similar constructions, this method represents an attractive solution for predicting the speaker's choice between three and more near-synonyms.

II. Previous research

Letting represents a subtype of causation. In Force Dynamics theory [8], letting is observed in situations when the Causer fails to override the Causee's intrinsic tendency towards some action or state. Unlike factitive causative constructions, such as *make* + V, *have* + V or the *into*-causative, the constructions of letting have been at the periphery of researchers' attention. However, there have been a few studies within a broader domain of infinitival complementation, such as [9] and [10], as well as some observations in general functionalist theories [11]. In particular, *let* is believed to express situations when the act of permission is construed as inseparable from the realization of the permitted event, while *allow* and *permit* only denote the prior condition for the permitted event [9]. In addition, it has been shown that *let* is more frequently used with the 1st and 2nd person matrix subjects than *allow* and *permit* and in the imperative form. In contrast, *allow* and *permit* occur more frequently with inanimate subjects [10].

The semantic difference is not the only factor that explains the use of the constructions. There are also social and collocational factors at play. For example, some grammars mention that the construction *permit* + *to* V is more formal than the construction with *allow* [12]. In addition, it has been observed that *let* forms a tight unit with some infinitives. Such expressions are synonymous with lexical causatives, e.g. *let fall* is similar to *drop* and *let know* is similar to *inform* [9].

The present paper aims to bring together these and other factors known from previous research of causation. These factors are operationalized as variables that help us predict the use of the three near-synonymous constructions in a large corpus.

III. Data and variables

1. Data

The data set is a sample from the British National Corpus (XML edition). To create the data set, we first extracted all forms of *let*, *allow* and *permit* that were followed by another verb within the context window of 6 words. Since *let* + V cannot be used in the passive (**He was let come*), only the active forms of the first verb were taken into account. Examples of adhortative let (e.g. *let's go*) were excluded. Next, a random sample of 882 instances for each construction was drawn (2646 examples in total), discarding all spurious hits. The examples were then coded for twenty variables, which are discussed in the following section. To speed up the coding process, we annotated the data set syntactically with the help of the Stanford Parser [13] and extracted the information about the main slot fillers of the constructional instances. All automatic annotations were manually checked.

2. Variables

1) Semantic variables

- *The semantic class of the Causer*, i.e. the entity that lets, allows or permits. This is a variable that can be represented as the animacy hierarchy [14], which is also known as the entrenchment hierarchy [15] or viewpoint/empathy hierarchy [16]. We used the classes from the hierarchy presented in (2):

  (2) Speaker > Hearer > Animate > Material (Physical) Object > Abstract

  Since there exist different versions of the hierarchy, this variable was coded as a categorical one and the classes were treated as unordered.
- *The semantic class of the Causee*, with the same classes as for the Causer.
- *Control of the Causee*, which shows whether the Causee has control over the permitted event or not. Controlling Causees are associated with less direct causation and weaker semantic integration of the causing and caused events than non-controlling ones, e.g. see [17].
- *Semantics of V2:* non-mental and mental.

2) Morphosyntactic variables

- *Tense, aspect and mood of V1:* imperative, Present Simple Indicative, Past Simple Indicative, Perfective, Progressive, Irrealis and Non-finite.
- *Valency of V2:* intransitive, transitive or passive.
- *Polarity:* positive or negative. The latter is operationalized as the presence of negative particles, pronouns or adverbs in the simple clause with the letting construction.
- *Coreferentiality:* the presence or absence of coreferentiality between the Causer and other participants of the causative situation.
- *Possession:* presence of absence of grammatical possession relationship between the Causer as the possessor and another participant as the possessee, formally marked by the possessive case or a possessive pronoun.

3) Social variables

- *Channel of communication*: written and spoken.
- *Domain of use:* public, educational, imaginary prose or other.

4) Collostructional measures

The collostructional measures are meant to represent the degree of association between each of the three letting constructions and the verbs that fill in the V2 slot, which are called collexemes. There exist a plethora of possible association measures for collexemes and constructions. For this study, we computed several popular measures that represent different aspects of relationships between a collexeme and a construction:

- *Attraction:* the proportion of collexeme X in the total frequency of construction A.
- *Reliance:* the proportion of occurrences of collexeme X in construction A.
- *Minimum Sensitivity:* in this context, the minimum score of Attraction and Reliance.
- *Collostructional strength:* a log-transformed *p*-value based on the Fisher exact test in collostructional analysis (see [18] for details).
- *ΔP* with verb as a cue, which represents the difference between the proportion of the verb in the total uses of the construction and the proportion of the same verb in the other constructions.
- *ΔP* with construction as a cue, which represents the difference between the proportion of the construction in the total frequency of the verb and the proportion

of the same construction in the total frequency of all other verbs.

These measures were computed for each instance of a letting construction with *let*, *allow* or *permit* observed in a given sentence and the corresponding V2. The verb frequencies were taken from a frequency list of lemmata based on the entire corpus. The constructional frequencies were computed with the help of a Python script, which counted all instances of *let*, *allow* and *permit* with a verbal complement in the syntactically parsed version of the corpus. In order to avoid multicollinearity, we decided to select one collostructional measure that would predict the use of the constructions the best. For this purpose, we fit several simple Bayesian multinomial mixed-effect regression models (see more details below) for each of these association measures. Having compared the models with the help of the Deviance Information Criterion (see Table 1), we concluded that the model with Minimum Sensitivity was the best. This variable was used for subsequent multivariate analyses presented in the next section.

TABLE I. DIC OF DIFFERENT COLLOSTRUCTIONAL MEASURES

Collostructional measure	**DIC**
Attraction	4595.74
Reliance	4957.77
Minimum Sensitivity	3582.81
Collostructional Strength	4159.99
ΔP with verb as a cue	4615.11
ΔP with construction as a cue	4959.13

5) *Formal variables*

- *formal linguistic distance*, which represents the formal distance in words between a verb of letting (V1) and the second predicate with or without *to* (V2). Words were defined as strings of alphabetic or numeric characters separated by white spaces.
- *Horror aequi*: the presence of another letting verb (*let*, *allow*, *permit* or *enable*) in the left context within the same sentence. *Horror aequi* is a tendency to avoid repetition of identical elements. The reason for considering this variable is to take into account the choice of a particular letting construction for stylistic purposes.
- *Length of V2 in characters*.

IV. A BAYESIAN MULTINOMIAL MIXED-EFFECT MODEL

We fit a Bayesian multinomial mixed-effect model with the letting construction as the response, the variables described above as fixed effects and the infinitives as random effects. The multinomial model contained two sets of coefficients, one where *allow* was compared with *let*, and the other where *permit* was compared with *let*.

The main advantage of the Bayesian GLM method is its flexibility. One can fit very complex models without running the risk of violating the assumptions that should be met in frequentist GLMs. A distinctive feature of Bayesian statistics is the use of the so called priors, i.e. the prior beliefs in the probability of some parameters. After the data are taken into account, the model returns the posterior probabilities of specific parameter values. These posterior probabilities depend on both the prior beliefs and the data, whereas the results of a frequentist model depend only on the data. In this model we used so-called flat priors. These priors have virtually no influence on the posteriors probabilities. However, when the data size is large, the choice of priors has hardly any effect.

In order to avoid strong autocorrelation in the Markov chain, our model was fit with a large number of iterations (310,000) and a thinning parameter of 100, which means that only every 100th iteration was taken into account, to reduce the effect of autocorrelation. We also used a burn-in period of 10,000 iterations, that is, removed the data based on the first 10,000 iterations in order to correct the initial sampling bias.

The estimates of the posterior probabilities (mean log-odds) were examined, along with the 95% Highest Density Intervals and the MCMC *p*-values. As an illustration, Figure 1 displays the effects of the semantic variables on the choice of *allow* and *permit* vs. *let* as the reference category. A manual check of all possible pairwise interactions between the variables has revealed a few significant cross-over interactions.

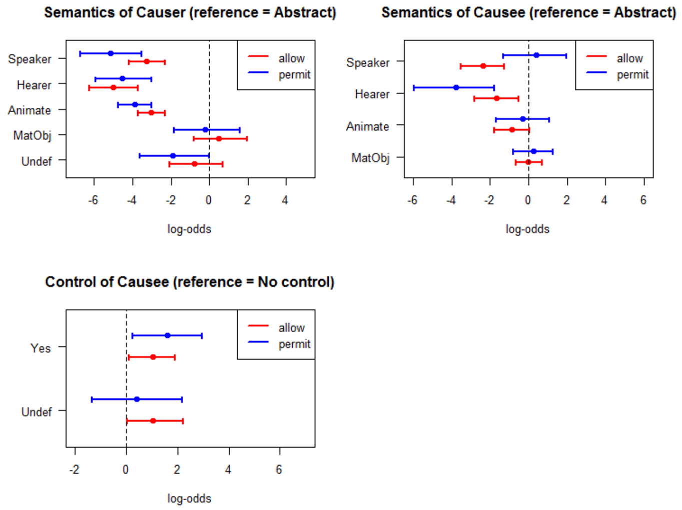

Fig. 1. Posterior mean log-odds and 95% Highest Density Intervals of three semantic variables.

To evaluate the goodness of fit, we computed the accuracy measure based a comparison between the predicted probabilities of *allow*, *let* and *permit* for every data point and the actual construction that was used in the given context. The accuracy, which is computed as the proportion of the correct predictions, was 70.7%. With the baseline at 33.3%, this is a clear improvement.

V. Results and discussion

The main result of the statistical analyses is a remarkable harmonic alignment of several literal and metaphoric distances in the sense that a smaller distance increases the odds of *let* and a greater distance increases the odds of *allow* and especially *permit*.

A. Formal linguistic distance between V1 and V2. The more words there are between the predicates, the higher the chances of *allow* (marginally significant) and *permit* (significant).

B. Conceptual distance between the causing and caused events expressed by V1 and V2. This distance is captured by the presence of the autonomous Causee who has control over the caused event. This feature increases the odds of *allow* and *permit* against *let*.

C. Cognitive distance between the speaker and the Causer and (to some extent) Causee on the animacy hierarchy, which has also been interpreted as the hierarchy of entrenchment, viewpoint or empathy. In general, the further from the speaker the participants are on this hierarchy, the higher the odds of *allow* and *permit* against *let*.

D. Communicative and social distance between the interlocutors. The odds of *allow* and *permit* increase when the communication is written, covers public topics (e.g. business, economy and politics) and does not involve immediate interaction between the speaker and the hearer. This distance is also mirrored in the length of the infinitives as an indicator of formality: the longer the infinitive, the higher the probability of *allow* and *permit*.

E. Collostructional distance, i.e. loose association between V1 and V2, as the inverse of collostructional fixation expressed by Minimum Sensitivity. The looser the association, the higher the chances of *allow* and *permit*.

In most situations, *permit* had more extreme posterior mean log-odds than *allow*. This means that *permit* more than *allow* differs from *let*. Moreover, although the variable *Horror aequi* had only marginal significance, *permit* seems to be used more often as a replacement for other letting verbs in order to avoid repetition.

Thus, we observe a remarkable alignment of different kinds of literal and metaphoric distances (or conversely, proximities). A crucial question is how to explain these results. There are at least two theories that can be useful for that purpose. One of them is iconicity theory. The alignment of the conceptual and formal distance can be explained by iconicity effects [17]. This principle can also account for the correlation between formal proximity and collostructional fixation. Moreover, iconicity can also explain the correspondence between social distance and length of linguistic forms [17], which manifests itself in this study in the length of V2 and in the type of the infinitive (bare or with the particle *to*).

However, iconicity theory has been recently shown to be outperformed by explanations that are based on usage [19]. Following this direction, it seems possible to explain more cases of alignment. First, highly salient Causers and Causees, which are high on the hierarchies of animacy, entrenchment, etc., may be the reason why *let* + V is the most frequent construction in informal conversations and spoken data, which, in its turn, may explain its shorter form (with a bare infinitive) in comparison with the two other constructions. The same can be said about the higher integration of events: more direct causation, on average, has higher salience than less direct causation. The differences in verbosity between more formal and less formal registers can explain the differences between the constructions with regard to the number of words between V1 and V2. Finally, high frequencies of some specific subschemata of *let* may also lead to a stronger collostructional association between the superordinate construction *let* + V and the corresponding V2 slot fillers.

All this suggests that the usage-based account can explain more cases of alignment and therefore might be considered superior to the account based on iconicity. Of course, the proposed explanation is only tentative, and more factual evidence of different types of alignment is needed in order to test and develop this theory.

References

[1] Gries, S. Th. 2003. *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement*. New York: Continuum.

[2] Heylen, K. 2005. A quantitative corpus study of German word order variation. In S. Kepser & M. Reis (eds.), *Linguistic evidence: Empirical, theoretical and computational perspectives*, 241–264. Berlin: Mouton de Gruyter.

[3] Arppe, A. 2008. Univariate, bivariate and multivariate methods in corpus-based lexicography – a study of synonymy. PhD diss., University of Helsinki.

[4] Divjak, D. 2010. *Structuring the Lexicon: a Clustered Model for Near-Synonymy*. Berlin: De Gruyter Mouton.

[5] Han, W., A. Arppe & J. Newman. In press. Topic marking in a Shanghainese corpus: from observation to prediction. To appear in Corpus Linguistics and Linguistic Theory.

[6] Stefanowitsch, A. & S. Th. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics 8*(2). 209–243.

[7] Hadfield, J. D. 2010. MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. *Journal of Statistical Software 33*(2). 1–22.

[8] Talmy, L. 2000. *Toward a Cognitive Semantics.* Cambridge: MIT Press.

[9] Duffley, P. J. 1992. *The English infinitive.* London: Longman.

[10] Egan, Th. 2008. *Non-finite Complementation: A usage-based study of infinitive and -ing clauses in English.* Amsterdam: Rodopi.

[11] Givón, T. 1980. The binding hierarchy and the typology of complements. *Studies in Language 4*(3). 333–377.

[12] Leech, G. & J. Svartvik. 1994. *A Communicative Grammar of English*. 2nd ed. London: Longman.

[13] Klein, D. & Ch. D. Manning. 2013. Accurate Unlexicalized Parsing. In *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics*.

[14] Silverstein, M. 1976. Hierarchy of Features and Ergativity. In R. M. W. Dixon (Ed.), *Grammatical Categories in Australian Languages*, 112–171. Canberra: Australian National University.

[15] Deane, P. D. 1992. *Grammar in Mind and Brain*. Berlin: Mouton de Gruyter.

[16] DeLancey, S. 1981. An interpretation of split ergativity and related patterns. *Language* 57(3): 626–657.

[17] Haiman, J. 1983. Iconic and economic motivation. *Language 59*(4). 781–819.

[18] Stefanowitsch, A. & S. Th. Gries. 2005. Covarying collexemes. *Corpus Linguistics and Linguistic Theory 1*(1). 1–43.

[19] Haspelmath, M. 2008. Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics* 19(1). 1–33.