Generating hypotheses for alternations at low and intermediate levels of schematicity. The use of Memory-Based Learning

According to usage-based linguistics, language variation addresses a functional need of the language user. That functional need may be dependent on the lexical realization of the varying constructions. For instance, while it may be useful to have an argument structure alternation express a particular semantic distinction for particular verbs or themes, that same distinction may be less relevant for other verbs or themes. As such, it has been argued that language variation should be investigated at low levels of schematicity, e.g. by studying argument structure alternations separately for various verbs, themes, etc. In this paper, we develop a data-driven procedure to do so, based on Memory-based Learning (MBL). The procedure focusses on generating hypotheses, is scalable, and can work with small datasets. It consists of three steps: (i) choosing features for the MBL classifier, (ii) running MBL analyses and selecting which analyses to put under further scrutiny, and (iii) inspecting which features were most useful in predicting the choice of variant in these analyses. Finally, the hypotheses that are inferred from these features are put to the test on separate data. As an example study, we investigate the Dutch *naar*-alternation.

Keywords: alternation, Memory-based Learning, data-driven, hypothesis generation, corpus

1 Introduction

One of the central assumptions of usage-based and variational linguistics is that language variation addresses a functional need of the language user (Geeraerts 2010: 263–265; Tagliamonte 2012: 1–2; Diessel 2017). This functional need may relate to e.g. organizing information structure (Jaeger 2010), or expressing semantic or social meaning (Speelman and Geeraerts 2009; Marzo et al. 2018). Meanwhile, corpus studies have traditionally focused on alternations between highly schematic constructions (Perek 2015: 105). These are constructions that contain no or only few fixed lexical elements, such as the English ditransitive and prepositional dative constructions (Bresnan et al. 2007), or the Dutch transitive and reflexive construction (Pijpops and Speelman 2017).

However, a functional need can in principle arise at any level of schematicity (cf. Diessel 2015: 207–209). For instance, Perek (2014) shows that the English *at*-alternation in (1) is used to express a difference in repetition for verbs of cutting, such as *chip*, *chisel* and *snip,* with the prepositional variant implying that the action is repeated. It may be useful to express this distinction for verbs of cutting, but less so for other verbs. Indeed, the same alternation expresses another distinction for verbs such as *kick* or *slap*, viz. whether or not contact with the target is entailed, as in (2). This alternation hence functions at a lower level of schematicity, as its determinants are dependent upon the lexical items in the constructions, in this case the verbs (Pijpops et al. 2021: 492–497).

(1) *Sam chipped **(at)** the rock*. (taken from Broccias 2001: 77)

(2) *He slapped **(at)** it with his other hand (…)* (taken from the British National Corpus, corpus-id: FS8-1809, cited in Perek 2015: 136)

Similarly, Boas (2010) and Röthlisberger et al. (2017: 700, 703) argue that in the study of the English dative alternation, meaning differences that are specific to particular lexical items may have been swept under the rug by pooling over various verbs. Put more generally, the determinants of an alternation may differ from one lexical item to the next, and researchers are therefore increasingly arguing that it is important

to study lexically-specific constructions at low and intermediate levels of schematicity (Croft 2003; Lehmann and Schneider 2012; Perek 2014: 83).

However, once this attention for lexically-specific constructions is put into practice, we run into three problems. First, we need hypotheses. If the meaning contrast expressed by an alternation may differ from one lexical item to the next, then it is hard to hypothesize beforehand what these contrasts might be. Second, as we investigate ever more concrete constructions, the number of distinct alternations ever increases. For instance, if we would have to investigate the Dutch dative alternation separately for each verb, we would have to investigate at least 252 distinct alternations (Colleman 2009: 597). If we would have to investigate it separately for each unique combination of a verb and a theme, that number of alternations would increase even further. Third, we will likely suffer from data scarcity. A dataset containing only the alternating instances of a single verb or a single verb-theme combination will necessarily be smaller than a dataset containing the instances of all alternating verbs. This is not a problem for highly frequent verbs, such as English *give*, but it can become an issue for less frequent lexical items.

We are thus in need of a method (i) that is data-driven, i.e. that focuses on generating hypotheses,; (ii) that is easily scalable, such that many concrete alternations can be investigated with relative ease; and (iii) that can work with limited amounts of data. In this paper, we propose to employ Memory-based Learning as such a technique (Daelemans and van den Bosch 2005). We will use Memory-Based Learning to investigate the Dutch *naar*-alternation as in (3). This is an alternation that occurs with 13 verbs in Dutch, viz. *bellen* 'ring', *graaien* 'grasp', *grabbelen* 'scramble', *grijpen* 'grab', *happen* 'bit', *jagen* 'hunt', *opbellen* 'ring up', *peilen* 'gauge', *schoppen* 'kick', *telefoneren* 'phone', *verlangen* 'desire', *vissen* 'fish' and *zoeken* 'search' (Pijpops 2019: 49–53). The data will be extracted from the Sonar corpus (Oostdijk et al. 2013a).

(3) *Technici zoeken nu **(naar)** de oorzaak van de rook.*
Technicians search now (to) the cause of the smoke
(Sonar-id: WS-U-E-A-0000205929.p.1.s.5)
'Technicians are now searching for the cause of the smoke.'

Section 2 introduces Memory-based Learning and Section 3 presents the data. Section 4 applies the technique, Section 5 evaluates the results and Section 6 summarizes the conclusions.

2 Using Memory-based Learning for hypothesis-generation

Memory-based Learning (MBL) is a *k*-nearest neighbor classifier that predicts the choice between linguistic variants for a new data point, based on a memory of previously observed data points that each represent a specific instance of variation. It does so by calculating the proximity of the new data point to all instances of its memory and then selecting the *k* training observations that most closely resemble the new data point. Based on these *k* observations, it finally predicts the variant used in the new data point. It thus does not build a model on the training data, such as a regression formula or a classification tree. Of course, the researcher still needs to specify the features with which the proximities between the new data point and the known observations are calculated. For other examples of MBL in fundamental linguistic research, see Scha et al. (1999), Keuleers and Daelemans (2007), Theijssen (2012), van den Bosch and Daelemans (2013), van den Bosch and Bresnan (2015), De Troij et al. (2021).

The following properties of MBL make it useful for our purposes. First, it is theoretically attractive, as it fits in well with exemplar-based cognitive theories of language proposed within usage-based linguistics (van den Bosch and Daelemans 2013). Second, it is conceptually simple, parsimonious in the number of

parameters that need to be set, and one can easily understand its behavior, compared to e.g. deep learning. For instance, in order to inspect why it makes a certain prediction for a new data point, one simply needs to look at the nearest neighbors. Third, MBL does not build a model on the training data. Hence, it makes no assumptions regarding the distribution of the training data; the only crucial assumption it does make, is that occurrences that are the most similar will tend to exhibit the same variant. As such, various MBL-analyses may be run automatically without needing to check model diagnostics – after all, there is no model.[1] In addition, a researcher can use whatever data is available, even if they are somewhat unbalanced. Fourth, MBL allows for various features to be tested simultaneously. Fifth, the categorical features may have a lot of levels. This is different from, e.g., regression analysis, where an important rule of thumb is not to include any more regressors than the number of observations of the least frequent response level divided by 20 (Speelman 2014: 530). As a result, it is usually not feasible to include e.g. the syntactic head of the agent or theme argument directly as a feature, since such a feature would have far too many levels and hence require too many regressors (although, see Van de Velde and Pijpops 2021 for a potential way to deal with this specific problem within regression analyses). Researchers would therefore usually a priori decide on more general categorizations such as the animacy or concreteness of the arguments (e.g. Pijpops and Speelman 2017; Röthlisberger et al. 2017). Of course, to decide on such categorizations, you already need to have hypotheses. By contrast, MBL does not require such high-level categorizations, and hence does not require a priori hypotheses. All of this means (i) that we can cast our nets wide when searching for hypotheses; (ii) that the MBL-analyses can easily be automated, so we can investigate a large number of lexically-specific alternations; and (iii) that Memory-based Learning can operate in data-scarce conditions (cf. van den Bosch and Bresnan 2015).

The data-driven procedure consists of the following three steps. In the first step, we choose the features that the MBL-classifier will use. In the second step, we run MBL-analyses separately for each verb or each unique verb-theme combination. When choosing which MBL-analyses to examine further, we will look at their predictive quality, quantified as C-indices. The C-index is equal to the Area Under the ROC Curve (AUC) for binary response variables such as the alternation studied here, and ranges in practice from 0.5 to 1. It is a measure of accuracy that is comparable across different baselines.[2] In the third step, we check the gain ratios of the features. These gain ratios indicate how useful each feature was for predicting the choice of variant (for its calculation, see Quinlan 1986). We then turn to the training data, and perform a qualitative analysis on the features with high gain ratios in order to determine what makes them so useful. Finally, we use this information to formulate a hypothesis about what distinction might drive the alternation.

We start with the first step. Two types of MBL-analysis are executed (cf. De Troij et al. 2021). The first is a window-based analysis that takes as features the five words to the left and the right of the start of the theme constituent, i.e. the place where the preposition appears when the prepositional variant is used. The preposition itself is of course not included in this window. The window thus may or may not include the syntactic head of the theme constituent, or the verb. All words are set in lower case and sentence boundaries are not crossed. If the sentence contains less than 5 words left or right from the start of the theme, the corresponding features have level *x*. The distinction between the explicit and implicit negation, e.g. *niet* 'not' vs. *geen* 'no', was removed from the features of the window-based analyses.

[1] Of course, it will still be interesting to investigate which features are useful for predicting the variant and why (see below), but this will be done to interpret the results, not to check whether any statistical assumptions were violated.

[2] Imagine a graph with the false positive rate on the x-axis and the true positive rate on the y-axis, and a line indicating the classifier's performance under different cut-offs, i.e. the receiver operator characteristics (ROC) curve. The AUC is then equal to the area under this line (Egan 1975; Daelemans et al. 2010: 34). For other ways of explaining the same measure, see Hosmer and Lemeshow (2000: 160–164) and Speelman (2014: 514–515).

Among prepositional constituents, the use of implicit negation is said to be only possible in certain regional varieties of Dutch or in contrastive contexts (Haeseryn et al. 1997: 1657-1658). This would mean that a choice for the prepositional variant would induce a preference for using explicit rather than implicit negation. A prediction of the variant based on the type of negation would therefore be circular.

The second type of MBL-analysis is parse-based. These analyses use information from the syntactic Alpino parses of the Sonar corpus (van Noord 2006). This type of analysis is only really feasible if automatic syntactic parsing is available. The features are listed below. The term AGENT refers the participant performing the action expressed by the verb. If the agent is not expressed, AGENT HEAD and AGENT TOPICALITY are coded as level *no agent*, and AGENT COMPLEXITY is 0. The term THEME refers to the participant with which the action expressed by the verb is concerned. We only used instances where the theme was expressed, because otherwise, neither of the variants was used.

- AGENT HEAD: word form of the syntactic head of the agent constituent, or *no agent*.
- AGENT TOPICALITY: *first person, second person, third person pronoun, definite noun, indefinite noun, subordinate clause, no agent* (Pijpops and Speelman 2017: 227–228).
- AGENT COMPLEXITY: natural logarithm of the number of words of the agent constituent (Pijpops et al. 2018: 524)).[3]
- VERB FORM: word form of the verb.
- THEME HEAD: word form of the syntactic head of the theme constituent. For the verb-theme combinations, this feature in effect reduces to the number of the theme head.
- THEME TOPICALITY: *definite, indefinite.*
- THEME COMPLEXITY: natural logarithm of the number of words of the theme constituent, not including the preposition *naar* 'to' if it is present (Pijpops et al. 2018: 524)).
- THEME-VERB ORDER: *theme-verb*, *verb-theme* (Pijpops et al. 2018: 533).

In addition, all analyses also take the features COUNTRY, with levels *Belgium* and *the Netherlands*, and TEXT TYPE, with a separate level for each component of the Sonar corpus (for a list, see Oostdijk et al. 2013b: 21). The components of the text messages, chat logs, tweets and discussion lists were not used, because the quality of the syntactic parses in these components was deemed too low (Oostdijk et al. 2013b: 49–50). The window-based analyses thus use 12 features, while the parse-based analyse use 10.

The advantage of the window-based analyses is that they cast their nets wider, i.e. they have the potential of detecting possible distinctions that are not captured by the parse-based features. Conversely, the parse-based analyses instantiate a more targeted search. In order to decide which MBL-analyses to put under further scrutiny, it will be useful to see which reach higher classification accuracy than the others, since these analyses are more likely to have picked up on some relevant distinction. While we are thus interested in the relative classification accuracy of the analyses, we are less interested in achieving the highest possible accuracy in absolute terms. Because of this, and because we want to keep the procedure technically manageable for linguists without much computational background, we have chosen not to run any parameter optimizing algorithms. Instead, we simply use parameter settings that are regarded as defaults for MBL, viz. the IB1-algorithm with the overlap metric and gain ratio feature weighting, *k* set to 5 and inverse linear class voting weights (Daelemans et al. 2010: 20–41). All presented analyses are the results of leave-one-out-testing (Weiss and Kulikowski 1991; Daelemans et al. 2010: 12, 40).

[3] A logarithmic transformation is used because constituent length appears to be processed in a logarithmic way by the human brain (Palliera et al. 2011: 2524).

In practice, this means that the classifier operates as follows. Given a 'target occurrence', it will predict whether the transitive or prepositional variant is used by calculating the distance from the target occurrence to all other occurrences. It does so according to Equation 1, where X is the target occurence, Y is another occurence, n is the number of features, w_i is the gain ratio of the feature at issue, which functions as a weight, x_i and y_i are the levels of the feature at issue of respectively X and Y, and $\delta(x_i, y_i)$ is the distance between x_i and y_i hat is calculated according to Equation 2.

$$\Delta(X, Y) = \sum_{i=1}^{n} w_i \, \delta(x_i, y_i)$$

Equation 1: Calculation of the distance between two occurrences

$$\delta(x_i, y_i) = \begin{cases} \left| \dfrac{x_i - y_i}{max_i - min_i} \right| & \text{if numeric, else} \\ 0 & \text{if } x_i = y_i \\ 1 & \text{if } x_i \neq y_i \end{cases}$$

Equation 2: Calculation of the distance between the same feature of two occurrences

Next, the classifier will select the 5 occurrences that are closest to the target occurrence, and have each of these 5 neighbors 'vote' for the transitive or prepositional variant. A neighbor always votes for the variant it appears in. These votes are then weighted as a function of the neighbor's distance to the target occurrence, with the weight calculated as in Equation 3, where d_j is the distance between the target occurrence and the neighbor, d_1 is the distance between the target occurrence and the nearest of the 5 neighbors, and d_5 is the distance between the target occurrence and the farthest of the 5 neighbors.

$$w_j = \begin{cases} \dfrac{d_5 - d_j}{d_5 - d_1} & \text{if } d_k \neq d_1 \\ 1 & \text{if } d_k = d_1 \end{cases}$$

Equation 3: Calculation of the voting weights

3 Data

All instances of the 13 alternating verbs that appeared with a theme were extracted from the Sonar corpus (Oostdijk et al. 2013a). We removed all instances for which the country of origin was unknown, and for which the theme was placed after the right bracket, in the *Nachfeld* of the clause, since such placement is not possible for the transitive variant (see Haeseryn et al. 1997: 1225–1400, Author 2018: 526). Next, the remainder were manually checked. All non-interchangeable instances were removed from the dataset, as per standard practice in alternation studies (cf. Colleman 2009: 599–601; Röthlisberger 2018: 53; Szmrecsanyi et al. 2016: 4–6, for a detailed overview of the selection, see Pijpops 2019: 141–147). These included instances that were extracted because of a parsing error, as well as a number of idiomatic expressions, such as *soort zoekt soort* 'birds of a feather flock together'. This left us with 93,668 instances.

Since *zoeken* 'search' is the most frequent verb by far, accounting for 65,774 instances, it will be investigated at the lowest feasible level of schematicity, that of unique verb-theme combinations. The other verbs are investigated at a slightly higher level of schematicity, viz. that of individual verbs. We require each verb and verb-theme combination to yield at least 40 interchangeable instances of each variant in our corpus to be put under scrutiny. Furthermore, we only look at verb-theme combinations with full nominal themes. This leaves 26 verb-theme combinations for *zoeken* 'search', from a total of 9070, and 6 verbs, viz. *bellen* 'ring', *grijpen* 'grab', *happen* 'bit', *peilen* 'gauge', *telefoneren* 'phone' and *verlangen* 'desire', to investigate.

Two of the retained verbs and verb-theme combinations are already studied in previous work, viz. *verlangen* 'desire', *peilen* 'gauge', *slachtoffer zoeken* 'search victim' and *woord zoeken* 'search word' (Pijpops 2019: 179–185, 193–196). The alternation for *verlangen* 'desire' was found to be determined by strong lexical biases of the themes, which indicated a distinction in construal, viz. between 'desire as demand' and 'desire as longing', the latter being associated with the prepositional variant. We hence expect a high gain ratio for the feature THEME HEAD. Meanwhile, the alternation for *peilen* 'gauge' exhibited a massive difference between the Belgian and Netherlandic varieties, and, albeit to a lesser extent than for *verlangen* 'desire', lexical biases of the themes. We hence expect a high gain ratio for COUNTRY, as well as, to a lesser extent, THEME HEAD.

For the *slachtoffer zoeken* 'search victim', it was shown that an aggressor searching for victims is predominantly expressed in the transitive variant, while a helper searching for victims is more often expressed in the prepositional variant. A high gain ratio for AGENT HEAD is thus expected. For *woord zoeken* 'search word', literally looking for specific words e.g. in a text is typically expressed in the transitive variant, whereas trying to come up with unspecific words when trying to explain something, is more often expressed in the prepositional variant. We hence expect the MBL-analyses to point towards this distinction. For the other verbs and verb-theme combinations, 30 randomly selected instances of each variant are be kept out of the analysis to later test the generated hypotheses.

4 Applying the procedure

We are now effectively left with 32 distinct alternations, namely the *naar*-alternation for 6 distinct verbs and 26 distinct verb-theme combinations. We create separate datasets for of all of these and run MBL-analyses on them. Since there are only 6 verbs to investigate, it is feasible to look at all of them. Figure 1 shows that their analyses all reach reasonably high C-indices. We ranked the verb-theme combinations of *zoeken* 'search' according to the C-index of their most successful type of MBL-analysis (i.e. window-based or parsed-based), and selected the top 10. These are the combinations for which the C-indices are shown in Figure 2. Figure 2 shows that the predictive performance of the two types of analyses may strongly diverge from one theme to the next, which can be interpreted as an indication that the factors driving the alternation may be rather different from one theme to the next.

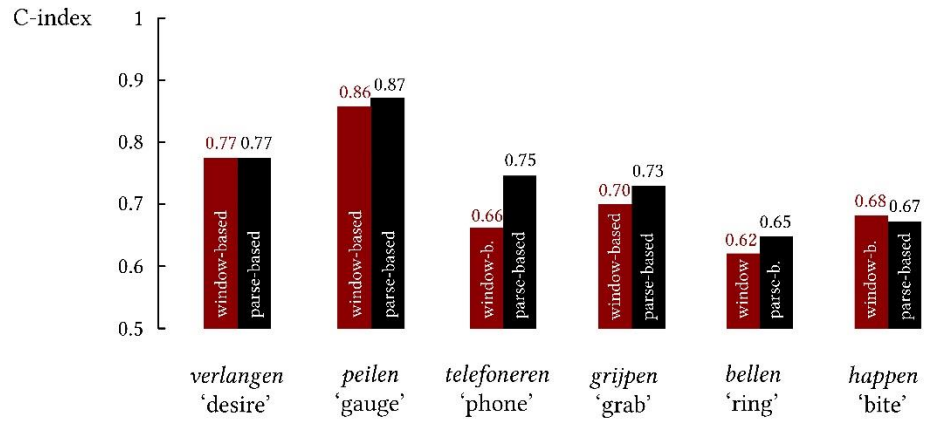


Figure 1: Predictive performance of the MBL-analyses for the verbs.

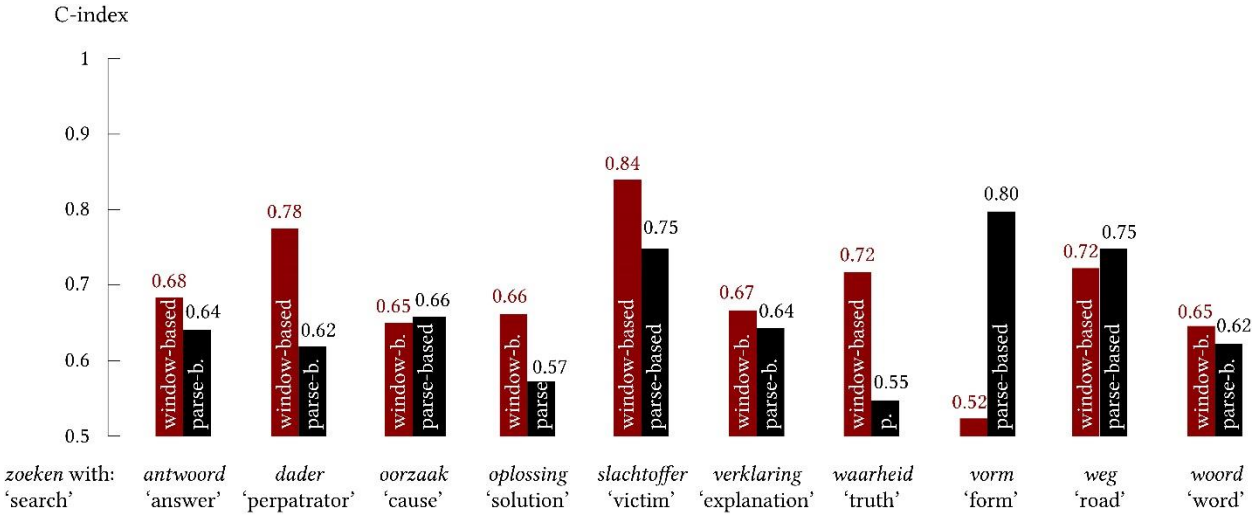
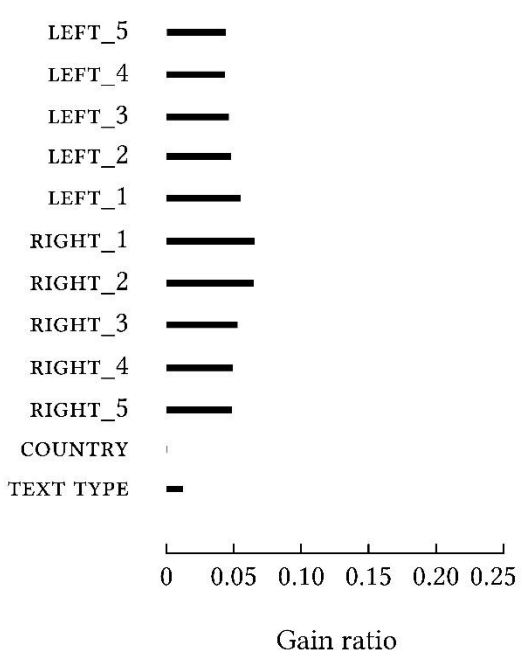


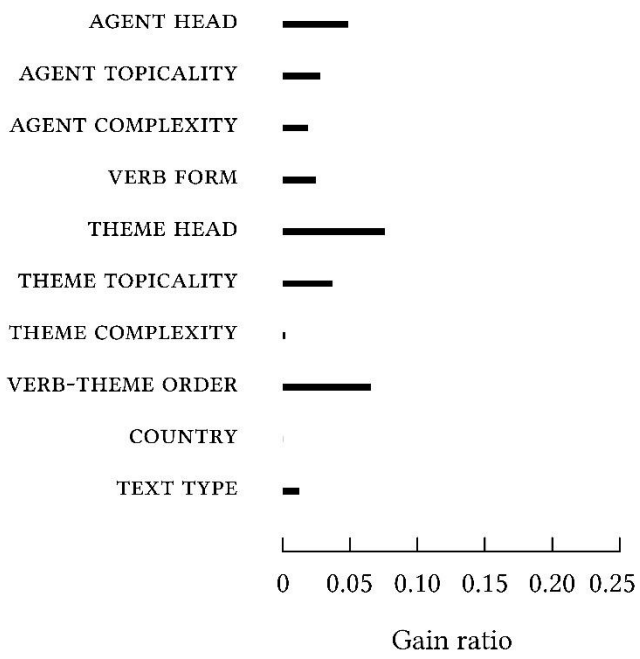
Figure 2: Predictive performance of the MBL-analyses for the verb-theme combinations.

Next, the gain ratios of these MBL-analyses are used to steer qualitative investigations of the data on which the MBL-analyses were run. Space restraints prevent us from discussing all qualitative analyses. Hence, only the qualitative analyses of one verb and two verb-theme combinations studied in previous work are discussed, viz. *verlangen* 'desire', *slachtoffer zoeken* 'seach victim' and *woord zoeken* 'search word', as well as those of one verb and two verb-theme combinations with the highest C-index, viz. *telefoneren* 'phone', *vorm zoeken* 'search form' and *weg zoeken* 'search road'. For the others, we simply list the results of the qualitative analyses, i.e. the generated hypotheses.

Figures 3-4 show the gain ratios for the verbs. The labels LEFT_5, LEFT_4,… in the graphs of the window-based analyses refer to the fifth, fourth,… word to the left. For *verlangen*, 'desire', we find that the theme head is indeed most useful for predicting the variant. We can then 'look under the hood', by inspecting which themes promote a prediction of the transitive variant or the prepositional variant in the training data. We find that e.g. *tegenprestatie* 'counter effort', *excuse* 'excuse', and *antwoord* 'answer' often occur with the transitive variant, while *dood* 'death', *huis* 'house' and *kind* 'child' often occur with the prepositional variant. This indeed points towards a distinction between 'desire as demand' and 'desire as longing'.

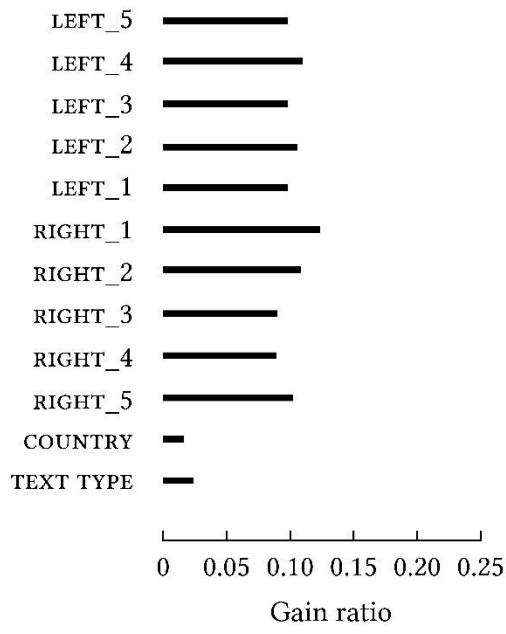


a. Window-based MBL-analyses

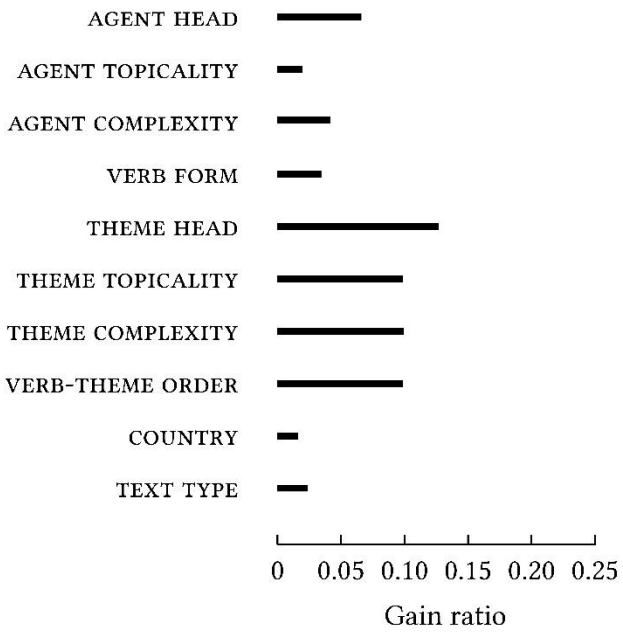


b. Parse-based MBL-analyses

Figure 3: Gain ratios for the verb *verlangen* 'desire'.



a. Window-based MBL-analyses



b. Parse-based MBL-analyses

Figure 4: Gain ratios for the verb *telefoneren* 'phone'.

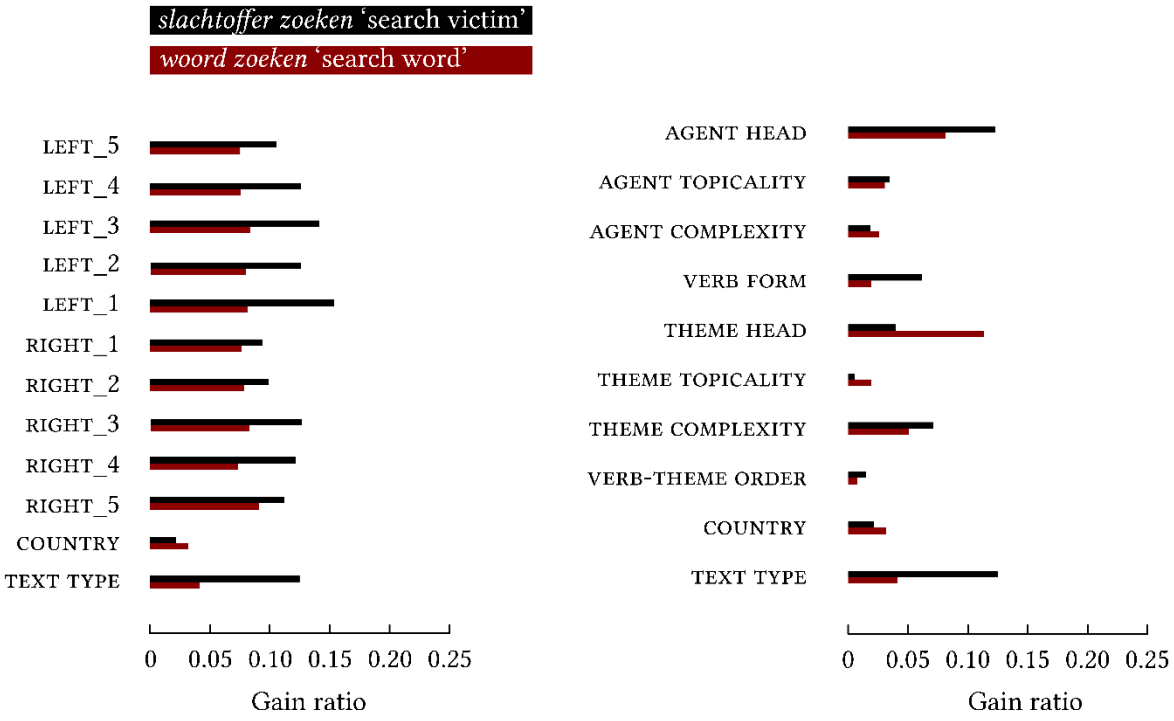
The gain ratios of *telefoneren* 'phone' show high values for the parse-based features relating to the theme, most notably THEME HEAD. When we look under the hood, it appears that human themes more often occur in the transitive variant, while non-human themes, i.e. collectives and inanimates, seem to prefer the prepositional variant. In the window-based analyses, we also find a little peak at RIGHT_1. This feature appears to point in the same direction of human vs. non-human themes. We hence formulate the following hypothesis for *telefoneren* 'phone': when the addressee is human, the transitive variant will be preferred, whereas when the addressee is not human, the prepositional variant will be more likely chosen.

The gain ratios for the four verb-theme-combinations can be found in Figures 5-6. For *slachtoffer zoeke*n 'search victim', the high gain ratio for AGENT HEAD was expected. Looking at which agents exhibit which preferences, we find a distinction between e.g. *zakkenrollers* 'pickpockets' and *daders* 'perpetrators' versus *reddingswerkers* 'rescue workers' and *duikers* 'divers'. When we look under the hood for TEXT TYPE, we find that the prepositional variant is more often used in the corpus component of the autocues: it appears that news readers more often talk about recent disasters where rescue workers are already looking for victims than about future crimes where criminals are still searching for victims. As for the feature LEFT_1, we find *brokstuk* 'wreckage', *speurhond* 'tracker dog' and *puin* 'rubble' to indicate the use of the prepositional variant. This again points towards the distinction between aggressors and helpers.

For *woord zoeken* 'search word', we find an notable peak for THEME HEAD: singular *woord* 'word' prefers the transitive variant, while plural *woorden* 'words' more often appears in the prepositional variant. From this, the relevant distinction could be inferred: literally looking for a word in a text usually involves just one word, while if a speaker means to express some proposition, multiple words are typically needed.

For *vorm zoeken* 'search shape', we find peaks in gain ratio for VERB-THEME ORDER, THEME TOPICALITY and AGENT HEAD. VERB-THEME ORDER shows a preference for the prepositional variant when the verb precedes the theme, which is consistent across the analyses (cf. Pijpops et al. 2018). Looking under the hood of THEME TOPICALITY and AGENT HEAD, we seem to find a distinction between sportspeople trying to get into their best condition, which evokes the use of the prepositional variant, and other instances of searching for forms. This will be our hypothesis for this combination.

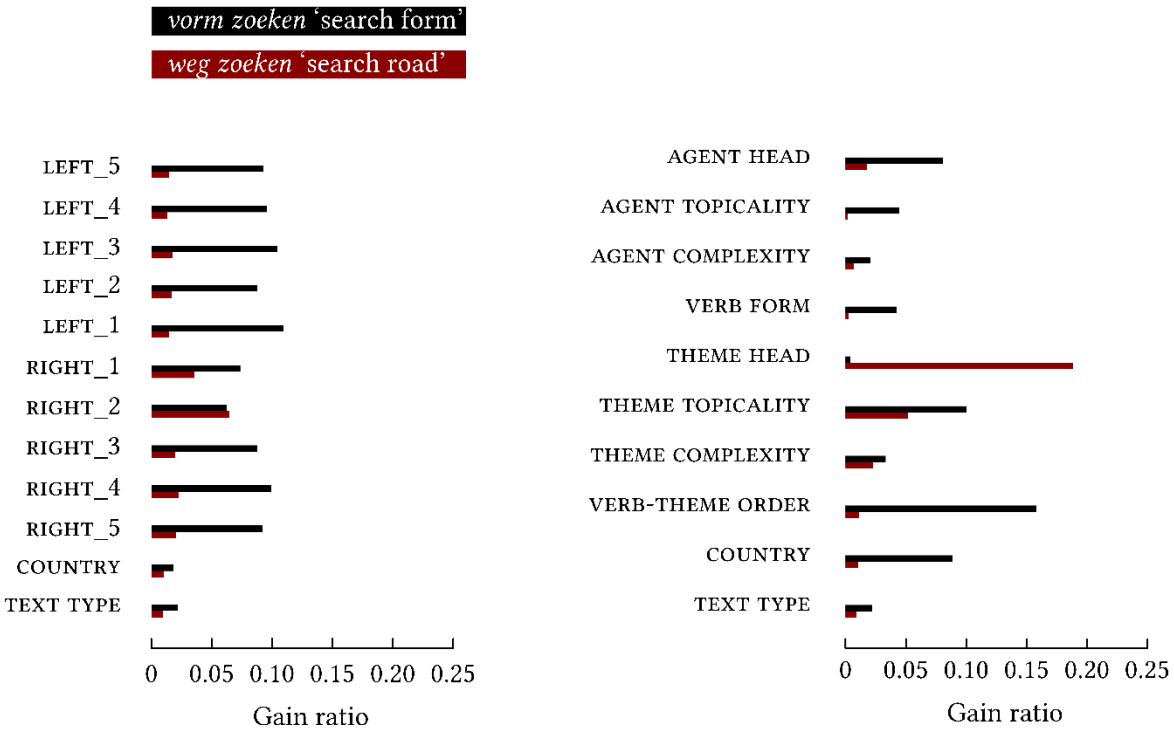
Lastly, for *weg zoeken* 'search road', the window-based analysis exhibits a marked peak in gain ratio for RIGHT_2 and a smaller one for RIGHT_1. For RIGHT_2, the words *eigen* 'own' and *weg* 'road' promote a choice for the prepositional variant, while *wegen* 'roads' and *om* 'for' prefer the prepositional variant. For RIGHT_1, we find possessive pronouns to be indicative of a choice for the transitive variant, whereas *wegen* 'roads' and *nieuwe* 'new' prefer the prepositional variant. For THEME HEAD, we again find singular *weg* 'roads' to promote the transitive variant, and plural *wegen* 'roads' or 'ways' the prepositional variant. Based on this, we formulate the hypothesis that when someone is finding their place in society or in a new job position or the like, this will be more often expressed in the transitive variant, while other forms of *weg zoeken* 'search road' will more often be expressed in the prepositional variant.



a. Window-based MBL-analyses

b. Parse-based MBL-analyses

Figure 5: Gain ratios for the verb-theme combinations *slachtoffer zoeken* 'search victim' and *woord zoeken* 'search word'.



a. Window-based MBL-analyses

b. Parse-based MBL-analyses

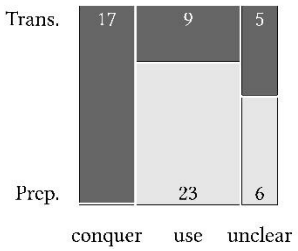
Figure 6: Gain ratios for the verb-theme combinations *vorm zoeken* 'search shape' and *weg zoeken* 'search road'.

The gain ratios of the other verbs and verb-theme combinations can be found in the Appendix. For *peilen* 'gauge', the results confirm our expectations with high values for COUNTRY and THEME HEAD. we also find high values for VERB-THEME ORDER and VERB FORM, which appear the indicate an effect of complexity (cf. Pijpops et al. 2018) '. For *grijpen* 'grab', we hypothesize that when *grijpen* can be translated as 'conquer', the transitive variant will be used, whereas when it involves a grabbing of a concrete object to be immediately used, the prepositional variant will be preferred. For *bellen* 'ring', we formulate the same hypothesis as for *telefoneren* 'phone'. For *happen* 'bite', our hypothesis states that when the biting succeeds, the transitive variant will be preferred.

For the combinations of *zoeken* 'search' with *antwoord* 'answer', *oorzaak* 'cause', *oplossing* 'solution' *verklaring* 'explanation' and *waarheid* 'truth', we hypothesize that if the answer, cause, etc. is already more or less known, but merely needs to be acquired or made into reality, then the transitive variant is employed, whereas if that is not the case, but an actual search still needs to be carried out, the prepositional variant will be preferred. To operationalize this distinction, we distinguish between instances where a locative adjunct already marks where the alternative, solution etc. needs to be sought, versus instances without a locative adjunct. For the combination *dader zoeken* 'search perpetrator', we hypothesize that when the authorities are searching for a perpetrator in a police investigation, there will be a preference for the prepositional variant.

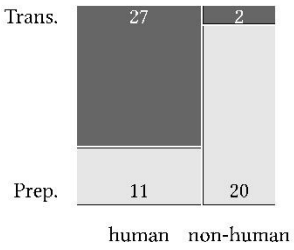
5 Evaluating the procedure

In the previous section, we have formulated a number of hypotheses based on the MBL-analyses. We can now evaluate the hypothesis-generating procedure by putting these hypotheses to the test. For the verbs and the verb-theme combinations that were studied in previous work, viz. *verlangen* 'desire', *peilen* 'gauge', *slachtoffer zoeken* 'search victim' and *woord zoeken* 'search word', the MBL-analyses did point towards the correct distinction. For the other verbs and verb-theme combinations, the 60 instances of each alternation that had been kept out of the MBL-analyses were manually annotated for the hypothesized distinctions, while blinded for the choice of variant. When an instance could not be clearly labelled as either of the hypothesized categories, it was labelled as *unclear*. The results are shown in mosaic plots in Figure 7. Mosaic plots are essentially bar charts, where the width of the columns is proportional to the number of observations in each category. For instance, Figure 7a shows that all 17 instances that were manually labelled to mean 'conquer' appeared in the transitive variant, while 9 of the instances labelled as 'use' appeared in the transitive variant, versus 23 in the prepositional variant. Meanwhile, 5 instances that were labelled 'unclear' appeared in the transitive variant, and 6 in the prepositional variant. The 'conquer' occurrences hence account for 28,3% of the data, the 'use' occurrences for 53,3% of the data, and the 'unclear' occurrences for 18,3% of the data. Therefore, the 'conquer' column takes up 28,3% of the width of the graph, the 'use' column 53,3% and the 'unclear' column 18,3%. Chi-squared tests or, where necessary, Fisher's exact tests are used to test for significance, with the *unclear* observations being excluded if there are any, and Cramer's V is an indication of effect size (Gries 2013: 183–186).



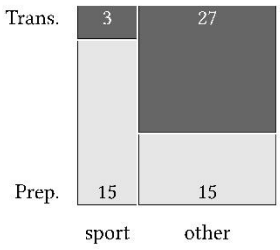
a. MEANING GRIJPEN for *grijpen* 'grab'

p < 0.0001
Cramer's V = 0.56



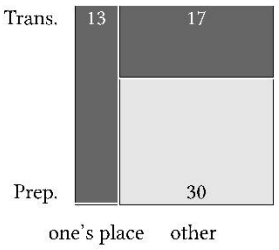
b. ADDRESSEE ANIMACY for *telefoneren* 'phone'

p < 0.0001
Cramer's V = 0.60



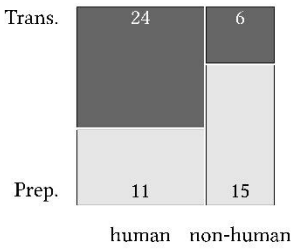
c. MEANING VORM for *vorm zoeken* 'search shape'

p = 0.0007
Cramer's V = 0.44



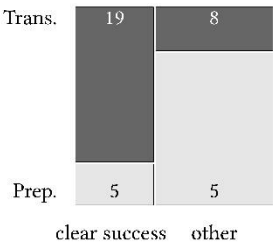
d. MEANING WEG for *weg zoeken* 'search road'

p < 0.0001
Cramer's V = 0.53



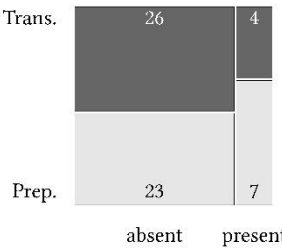
e. ADDRESSEE ANIMACY for *bellen* 'phone'

p = 0.0149
Cramer's V = 0.31



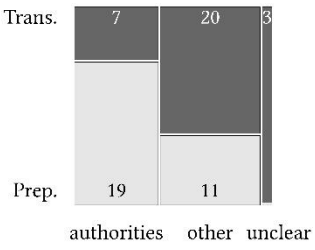
f. CONATION for *happen* 'bite'

p < 0.0001
Cramer's V = 0.56



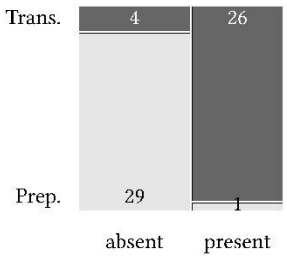
g. LOCATIVE ADJUNCT for *antwoord zoeken* 'search answer'

p = 0.3169
Cramer's V = 0.13



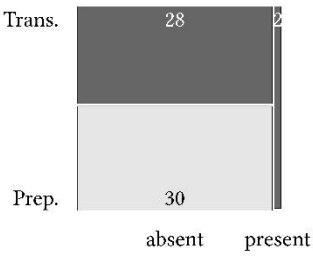
h. AGENT TYPE for *dader zoeken* 'search perpetrator'

p = 0.0028
Cramer's V = 0.43



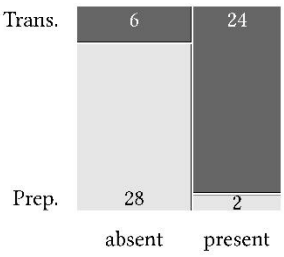
i. LOCATIVE ADJUNCT for *oorzaak zoeken* 'search cause'

p < 0.0001
Cramer's V = 0.84



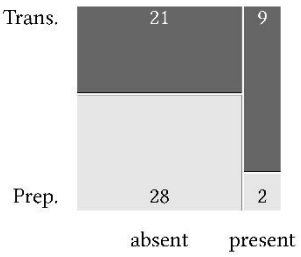
j. LOCATIVE ADJUNCT for *oplossing zoeken* 'search solution'

p = 0.4915
Cramer's V = 0.19



k. LOCATIVE ADJUNCT for *verklaring zoeken* 'search explanation'

p < 0.0001
Cramer's V = 0.74



l. LOCATIVE ADJUNCT for *waarheid zoeken* 'search truth'

p = 0.0195
Cramer's V = 0.30

Figure 7: Mosaic plots of the hypothesis testing analyses.

All hypotheses are confirmed by the tests in Figure 7, except those of *antwoord zoeken* 'search answer' and *oplossing zoeken* 'search solution'. Including the analyses for *verlangen* 'desire', *peilen* 'gauge',

slachtoffer zoeken 'search victim' and *woord zoeken* 'search word', the procedure thus pointed towards the expected or a confirmed hypothesis 14 out of 16 times. Furthermore, in the case of *oplossing zoeken* 'search solution', the failure to confirm the hypothesis may well be due to a lack of data. The mosaic plot does show the hypothesized tendency – a preference for the transitive variant when locative adjuncts are present – but our testing data happened to contain only two instances with a locative adjunct. That is simple not sufficient either to confirm or refute the hypothesis. In sum, we can therefore evaluate the hypothesis-generating procedure to be generally successful.

How to interpret these results for the *naar*-alternation? It appears that the alternation functions at a fairly low level of schematicity. The meaning contrast that is expressed by the alternation is dependent upon the verb in question, with some (near-)synonymous verbs, such as *bellen* 'phone' and *telefoneren* 'phone', clustering together. Such a situation can also be found for the English *at*-alternation (Perek 2015: 105–144). Meanwhile, for the highly frequent verb *zoeken* 'search', we observe a similar situation at an even lower level of schematicity. The meaning contrast expressed by the alternation appears to be dependent upon the theme in question, with semantically related themes clustering together. For instance, we have found the same contrast to be at play for *antwoord* 'answer', *oorzaak* 'cause', *oplossing* 'solution', *verklaring* 'explanation', and *waarheid* 'truth', while another contrast was confirmed for *dader* 'perpetrator' and *slachtoffer* 'victim'.

6 Conclusions

If functional needs can arise at any level of schematicity, then we should expect the determinants of language variation to also operate at any level of schematicity. Put concretely, language users may use an alternation to express a particular semantic distinction that is highly relevant in one particular lexical context, while in a different lexical context, it may be more useful to express another semantic distinction. This paper proposed a hypothesis-generating procedure to track down such lexically-specific determinants of language variation. The procedure can handle diverse types of features, each with many levels, can work with limited amounts of data, and is scalable. To end this article, the three steps of the hypothesis generating procedure are repeated below. After these steps, one would probably want to test the generated hypotheses on unseen data.

1. Choose the features that the MBL-classifier will use and annotate your data for them – preferably automatically, for instance by simply selecting the *n* words to the left and the right of the variant.
2. Run MBL-analyses for various subsets of the data, e.g. for various lexemes or combinations of lexemes, and use the C-indices to decide which analyses to put under further scrutiny.
3. Use the gain ratios of these analyses to decide which features to investigate further, and interpret these features to formulate hypotheses.

References

Boas, Hans. 2010. The syntax-lexicon continuum in Construction Grammar. A case study of English communication verbs. *Belgian Journal of Linguistics* 24(1). 54–82.

Bosch, Antal van den and Joan Bresnan. 2015. Modeling dative alternations of individual children. *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning*, 103–112. Lisbon, Portugal: Association for Computational Linguistics.

Bosch, Antal van den and Walter Daelemans. 2013. Implicit Schemata and Categories in Memory-based

Language Processing. *Language and Speech* 56(3). 309–328.
Bresnan, Joan, Anna Cueni, Tatiana Nikitina and Rolf Harald Baayen. 2007. Predicting the dative alternation. In Gerolf Bouma, Irene Krämer & Joost Zwarts (eds.), *Cognitive Foundations of Interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.
Broccias, Cristiano. 2001. Allative and ablative at-constructions. In Mary Adronis, Christopher Ball, Elston Heide & Sylvain Neuvel (eds.), *CLS 37: The Main Session. Papers from the 37th Meeting of the Chicago Linguistic Society*, 67–82. Chicago: Chicago Linguistic Society.
Colleman, Timothy. 2009. Verb disposition in argument structure alternations: a corpus study of the dative alternation in Dutch. *Language Sciences* 31(5). 593–611.
Croft, William. 2003. Lexical rules vs. constructions. A false dichotomy. In Hubert Cuyckens, Thomas Berg, René Dirven & Klaus-Uwe Panther (eds.), *Motivation in language: studies in honor of Günter Radden*, 49–68. Stanford: CSLI Publications.
Daelemans, Walter and Antal van den Bosch. 2005. *Memory-based language processing*. Cambridge: Cambridge University Press.
Daelemans, Walter, Jakub Zavrel, Ko van der Sloot and Antal van den Bosch. 2010. TiMBL: Tilburg Memory-Based Learner Reference Guide. Tilburg.
De Troij, Robbert, Stefan Grondelaers, Dirk Speelman and Antal van den Bosch. 2021. Lexicon or grammar? Using memory-based learning to investigate the syntactic relationship between Netherlandic and Belgian Dutch. *Natural Language Engineering*. 1–19.
Diessel, Holger. 2015. Usage-based construction grammar. In Ewa Dąbrowska & Dagmar Divjak (eds.), *Handboek of Cognitive Linguistics*, 296–322. Berlin: De Gruyter Mouton.
Diessel, Holger. 2017. Usage-Based Linguistics. In Mark Aronoff (ed.), *Oxford Research Encyclopedia of Linguistics*. Oxford: Oxford University Press.
Egan, James. 1975. *Signal detection theory and ROC analysis*. (Academic Press Series in Cognition and Perception). New York: Academic press.
Geeraerts, Dirk. 2010. *Ten Lectures on Cognitive Sociolinguistics*. Beijing: Beijing Foreign language teaching and research press.
Gries, Stefan Thomas. 2013. *Statistics for linguistics with R. A practical introduction*. 2nd edn. Berlin: De Gruyter.
Haeseryn, Walter, Kirsten Romijn, Guido Geerts, Jaap de Rooij and Maarten van den Toorn. 1997. *Algemene Nederlandse Spraakkunst*. Groningen: Nijhoff.
Hosmer, David and Stanley Lemeshow. 2000. *Applied logistic regression*. 2nd edn. New York: Wiley.
Jaeger, Florian Tim. 2010. Redundancy and Reduction: Speakers Manage Syntactic Information Density. *Cognitive Psychology* 61(1). 23–62.
Keuleers, Emmanuel and Walter Daelemans. 2007. Memory-based learning models of inflectional morphology: A methodological case study. *Lingue e Linguaggio* 6(2). 151–174.
Lehmann, Hans Martin and Gerold Schneider. 2012. Syntactic variation and lexical preference in the dative-shift alternation. In Joybrato Mukherjee & Magnus Huber (eds.), *Corpus Linguistics and Variation in English*, 65–75. Amsterdam: Rodopi.
Marzo, Stefania, Eline Zenner and Dorien Van De Mieroop. 2018. When sociolinguistics and prototype analysis meet: The social meaning of sibilant palatalization in a Flemish Urban Vernacular. In Eline Zenner, Ad Backus & Esme Winter-Froemel (eds.), *Cognitive Contact Linguistics Placing usage, meaning and mind at the core of contact-induced variation and change.*, 127–156. Berlin: Mouton De Gruyter.
Noord, Gertjan van. 2006. At Last Parsing Is Now Operational. In Piet Mertens, Cédric Fairon, Anne Dister & Patrick Watrin (eds.), *TALN 2006. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, 20–42. Louvain-la-Neuve: Cental.
Oostdijk, Nelleke, Martin Reynaert, Véronique Hoste and Ineke Schuurman. 2013a. The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch. In Peter Spyns & Jan Odijk (eds.), *Essential Speech and Language Technology for Dutch, Theory and Applications of Natural Language Processing*, 219–247. Heidelberg: Springer.
Oostdijk, Nelleke, Martin Reynaert, Véronique Hoste and Ineke Schuurman. 2013b. *SoNaR User Documentation*.

Palliera, Christophe, Anne-Dominique Devauchellea and Stanislas Dehaenea. 2011. Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences - PNAS* 108(6). (From the Cover). WASHINGTON: National Academy of Sciences. 2522–2527.

Perek, Florent. 2014. Rethinking constructional polysemy: The case of the English conative construction. In Dylan Glynn & Jus Robinson (eds.), *Polysemy and synonymy: Corpus methods and applications in cognitive linguistics*, 61–85. Amsterdam/Philadelphia: John Benjamins.

Perek, Florent. 2015. *Argument structure in usage-based construction grammar: experimental and corpus-based perspectives*. Amsterdam/Philadelphia: John Benjamins.

Pijpops, Dirk. 2019. How, why and where does argument structure vary? A usage-based investigation into the Dutch transitive-prepositional alternation. Dissertation University of Leuven.

Pijpops, Dirk and Dirk Speelman. 2017. Alternating argument constructions of Dutch psychological verbs. A theory-driven corpus investigation. *Folia Linguistica* 51(1). 207–251.

Pijpops, Dirk, Dirk Speelman, Stefan Grondelaers and Freek Van de Velde. 2018. Comparing explanations for the Complexity Principle. Evidence from argument realization. *Language and Cognition* 10(3). 514–543.

Pijpops, Dirk, Dirk Speelman, Stefan Grondelaers and Freek Van de Velde. 2021. Incorporating the multi-level nature of the constructicon into hypothesis testing. *Cognitive Linguistics* 32(3). 487–528.

Quinlan, John Ross. 1986. Induction of decision trees. *Machine Learning* 1(1). 81–106.

Röthlisberger, Melanie. 2018. Regional variation in probabilistic grammars: A multifactorial study of the English dative alternation. Dissertation University of Leuven.

Röthlisberger, Melanie, Jason Grafmiller and Benedikt Szmrecsanyi. 2017. Cognitive indigenization effects in the English dative alternation. *Cognitive Linguistics* 28(4). 673–710.

Scha, Renko, Rens Bod and Khalil Sima'an. 1999. A memory-based model of syntactic analysis: data-oriented parsing. *Journal Of Experimental & Theoretical Artificial Intelligence* 11(3). 409–440.

Speelman, Dirk. 2014. Logistic regression: A confirmatory technique for comparisons in corpus linguistics. In Dylan Glynn & Justyna A. Robinson (eds.), *Corpus Methods for Semantics: Quantitative studies in polysemy and synonymy*, 487–533. Amsterdam: John Benjamins.

Speelman, Dirk and Dirk Geeraerts. 2009. Causes for causatives: the case of Dutch "doen" and "laten." In Ted Sanders & Eve Sweetser (eds.), *Causal Categories in Discourse and Cognition*, 173–204. Berlin: Mouton de Gruyter.

Szmrecsanyi, Benedikt, Douglas Biber, Jesse Egbert and Karlien Franco. 2016. Toward more accountability: Modeling ternary genitive variation in Late Modern English. *Language Variation and Change* 28(1). 1–29.

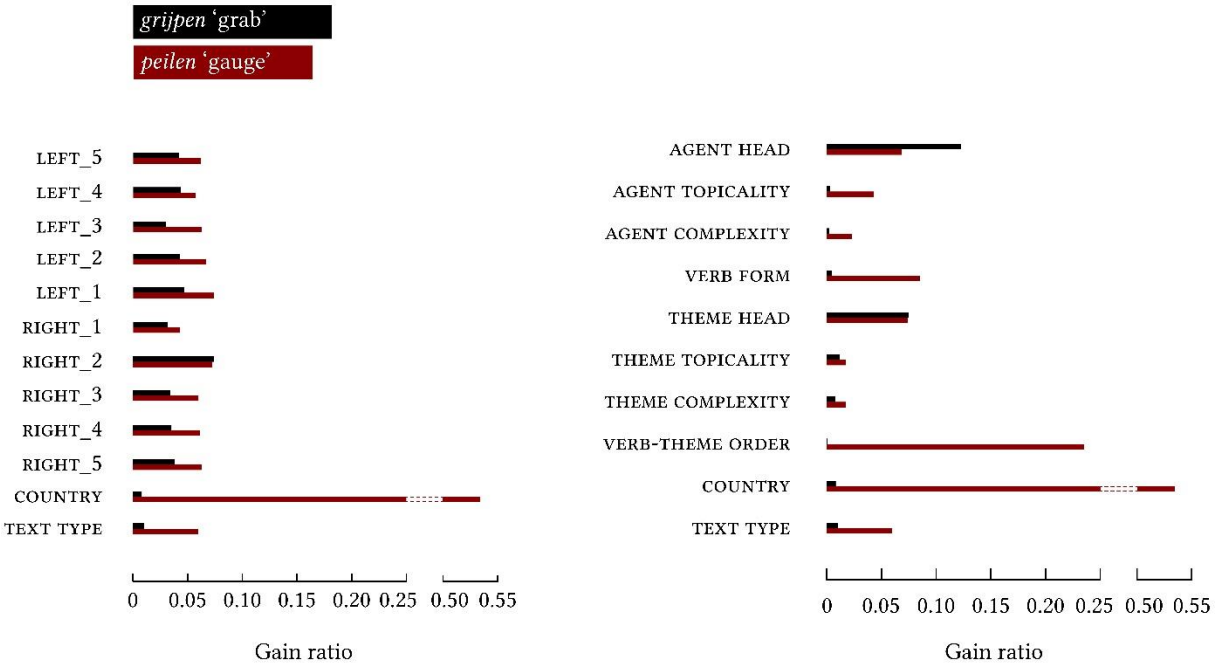
Tagliamonte, Sali. 2012. *Variationist sociolinguistics: change, observation, interpretation*. (Language in Society 40). Chichester: Wiley-Blackwell.

Theijssen, Daphne. 2012. Making Choices. Modelling the English dative alternation. Dissertation Radboud University Nijmegen.

Van de Velde, Freek and Dirk Pijpops. 2021. Investigating Lexical Effects in Syntax with Regularized Regression (Lasso). *Journal of Research Design and Statistics in Linguistics and Communication Science* 6(2). 166–199.

Weiss, Sholom and Casimir Kulikowski. 1991. *Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems*. San Mateo: Kaufmann.

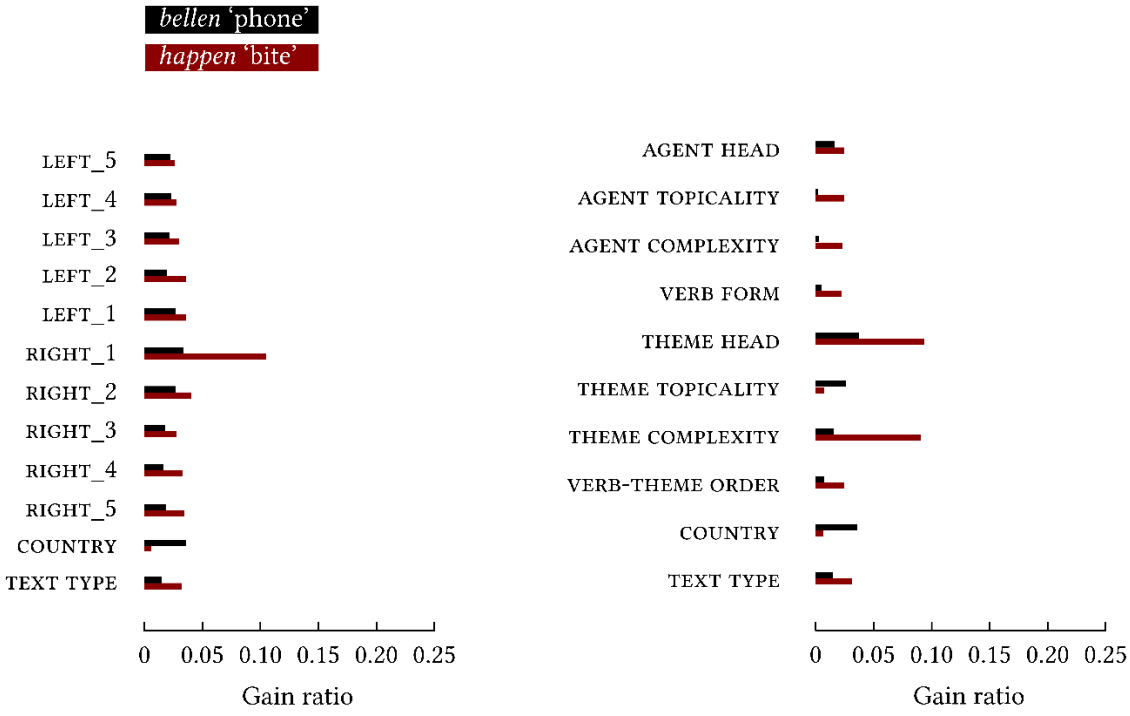
Appendix



a. Window-based MBL-analyses

b. Parse-based MBL-analyses

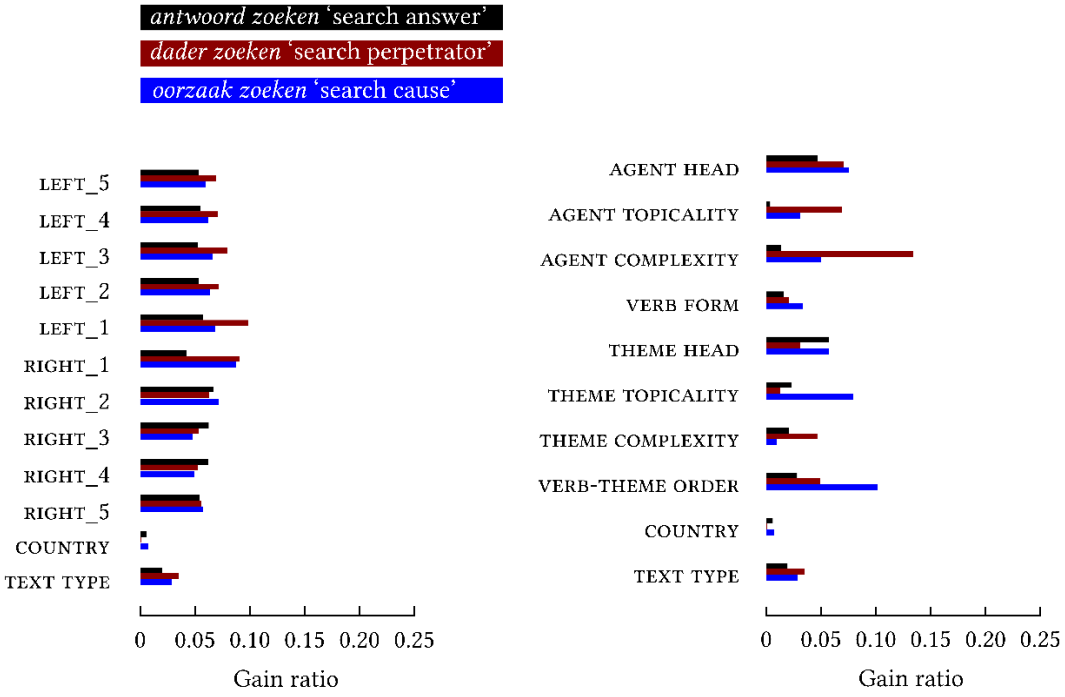
Figure 9: Gain ratios for the verbs *grijpen* 'grab and *peilen* 'gauge'.



c. Window-based MBL-analyses

d. Parse-based MBL-analyses

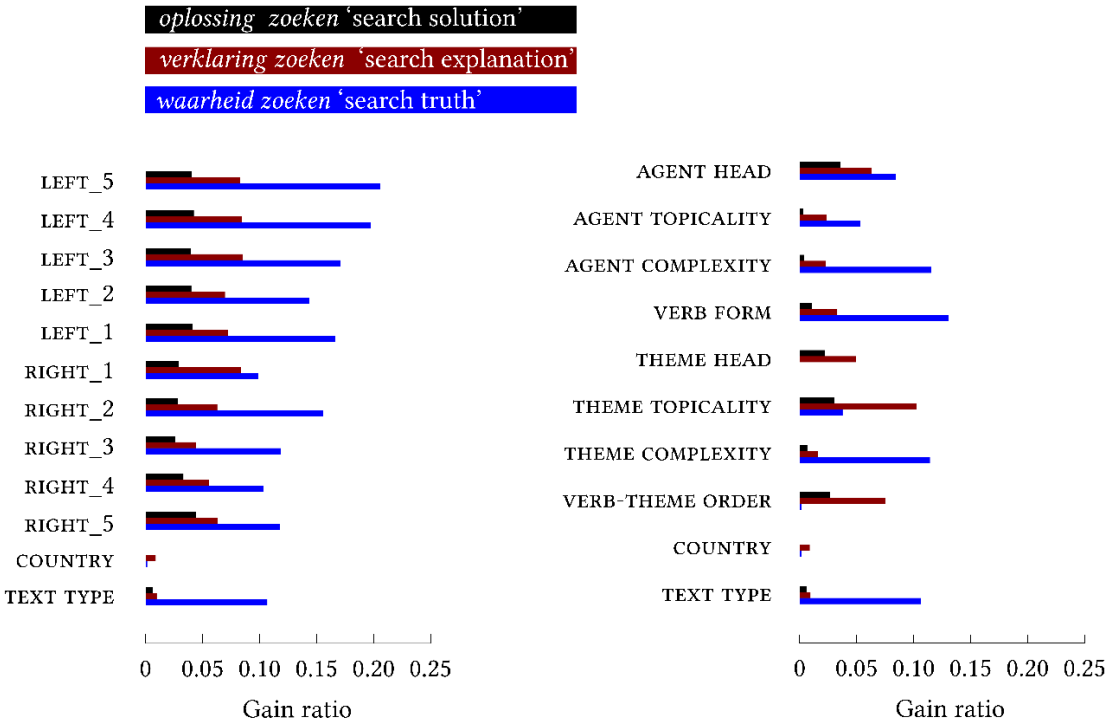
Figure 9: Gain ratios for the verbs *bellen* 'phone' and *happen* 'bite'.



a. Window-based MBL-analyses

b. Parse-based MBL-analyses

Figure 10: Gain ratios for the verb-theme combinations *antwoord zoeken* 'search answer', *dader zoeken* 'search perpetrator' and *oorzaak zoeken* 'search cause'.



a. Window-based MBL-analyses

b. Parse-based MBL-analyses

Figure 11: Gain ratios for the verb-theme combinations *oplossing zoeken* 'search solution' and *verklaring zoeken* 'search explanation' and *waarheid zoeken* 'search truth'.