Journal of Research Design and Statistics in Linguistics and Communication Science

JRDS (PRINT) ISSN 2052-417X
JRDS (ONLINE) ISSN 2052-4188

Article

Investigating Lexical Effects in Syntax with Regularized Regression (Lasso)

Freek Van de Velde and Dirk Pijpops

Abstract

Within usage-based theory, notably in construction grammar though also elsewhere, the role of the lexicon and of lexically-specific patterns in morphosyntax is well recognized. The methodology, however, is not always sufficiently suited to get at the details, as lexical effects are difficult to study under what are currently the standard methods for investigating grammar empirically. In this short article, we propose a method from machine learning: regularized regression (Lasso) with k-fold cross-validation, and compare its performance with a Distinctive Collexeme Analysis.

KEYWORDS: CROSS-VALIDATION; REGULARIZATION; LASSO; MACHINE LEARNING; CORPUS LINGUISTICS; COLLOSTRUCTIONAL ANALYSIS; DISTINCTIVE COLLEXEME ANALYSIS; OVERFITTING

1. Introduction

In usage-based construction grammar, there is no clear distinction between the lexicon and syntax, a view often succinctly summarized in Goldberg's (2006: 18) famous dictum 'It's constructions all the way down'. Argument constructions and syntactic frames have their own meaning. Some of these constructions have overlapping meanings and can combine with the same verbs or verb constructions: the double object constructions and the prepositional object construction can both express transfer of possession, and can occur with verbs like *give*, *donate*, *pass*, *transfer*, *send*, etc. The meaning of these two argument

Affiliation

KU Leuven, Belgium
Email: freek.vandevelde@kuleuven.be (corresponding author)

Université de Liège

JRDS VOL 6.2 2019 166–199
https://doi.org/10.1558/jrds.18964

constructions is sufficiently related to study them in combination, treating them as an 'alternation', in this case the dative alternance (see Colleman, 2006; Röthlisberger, 2018; Zehentner, 2019; De Vaere, 2020 for book-length treatments), or even as an 'allostruction' (Cappelle, 2006).

The choice between the double object construction and the prepositional object construction, also known as the dative alternance, is driven multifactorially (Bresnan *et al.*, 2007 and the references mentioned above), by animacy and topicality among other factors, but at the same time, not all verbs that engage in this argument construction choice have the same weights for these factors. Some of these verbs are more sensitive to these factors than others. Moreover some verbs may have a very strong predilection for one of the variants. This is probably true for all lexically underspecified constructions.

Unlike lexicalist or projectionist approaches like e.g. Rappaport-Hovav and Levin (2008), usage-based construction grammar does not assume that the verb meaning determines the construction it combines with. Why the distribution of the constructions can be skewed for different verbs is not straightforwardly explained, and has to do with frequency of use, entrenchment, analogy and partially also with lexical semantics (see Perek, 2015; Diessel, 2019: Ch. 7; Pijpops, 2019).

To dig deeper into this interesting issue, usage-based construction grammar and related approaches need to integrate the lexical semantic effects into the multifactorial accounts they put forth. The preferred method for investigating multifactorially-driven phenomena like argument realization alternations, has, for the last quarter century, been generalized linear regression, mostly with the logit link (see Gries, 2000; Grondelaers, 2000 for pioneering work, and Speelman, 2014 for a good, short introduction). This approach is attractive because it can deal with a multifactorial design combining both extralinguistic predictors (such as age, gender, and socio-economic status), and intralinguistic predictors, and because it gives both effect size and significance. Moreover, it is a versatile technique because it can simultaneously integrate numeric and categorical variables, and interaction effects. The method yields good results: the models achieve high levels of explained variance, as measured by R^2, and have been validated on experimental results (Bresnan and Ford, 2010; Klavan and Divjak, 2016). In its simplest form, the method has a hard time entering the lexical effects of, for instance, the different verbs, into the multifactorial design. In principle, one could enter all the different verbs as levels of a categorial predictor in the analysis, but most datasets are too small to cope with the inflation of the predictor set, even if the verb factor would be treated as a main effect, ignoring interactions. Several solutions have been adopted to dig into the lexical effects, but none are without drawbacks:

1. Building the statistical model for one verb only. In the dative alternation, one could look solely at the prototypical example of *give* in its transfer of possession sense. The advantage is obvious: the lexical effect is kept under tight control. But this comes at the cost of generalizability: the result is robust, but cannot be straightforwardly extrapolated to other verbs (see also Röthlisberger *et al.*, 2017: 700, 703).
2. Building a mixed-effect model with different intercepts, and possibly slopes, for the different verbs. This is the state-of-the-art solution in corpus linguistics today (Gries, 2015). There are disadvantages, however, and they are not always recognized. It is common practice to prioritize the random effects, in order to be conservative about the effect of the focus predictors in the fixed effects. Avoiding Type I errors, however, comes at the cost of an increase in Type II errors: the amount of variance that is left to be explained by the fixed effects may be too small to distinguish subtle effects. This may be the case if there is collinearity between the effects of different verbs and the focal variables. Corpus linguistics – to the extent that they have adopted the mixed-model approach – also tend to follow the 'keep it maximal' credo, adding random slopes to their models, even in cases where this is not fully warranted, further reducing the power of the fixed effects (see Winter, 2020: 242). In the mixed-model approach, the coefficients for the predictors are then ~~averaged out over~~ different lexemes that occur in the construction, effectively obscuring the precise lexical effects. A random intercept for verbs in a constructional alternation does only add one term to the model, by estimating the variation around the overall intercept, not an additional term for each verb (see Winter, 2020: 238). The common strategy to investigate the best linear unbiased predictors (BLUPS) for the by-word adjustments, as a way to look into lexical differences arguably stretches the purpose of random effects, which are meant to model the association structure in the data, with the fixed effects modeling systematic trends (see Van de Velde *et al.*, forthcoming). If you are really interested in the effects of the individual verbs, why are they not part of the fixed-effect structure, where they belong as focal variables? If you prioritize the random-effect structure over the fixed-effect structure in the model assembly to avoid Type I errors (the common practice, as pointed out above), you increase the Type I errors via the back door, if you then use the BLUPS to investigate the lexical differences. Moreover, the corpus will yield many more observations from frequent verbs than from infrequent verbs. The typically Zipfian frequency distribution of verbs will display a long tail of sparsely attested levels. The maximum likelihood estimation (typically

by Laplace approximation) might have a hard time converging on an adequate model: the size of the random intercepts – let alone slopes – may not be reliably estimable with underpopulated levels of the random factors. An often used 'solution' is to bin all verbs with few observations (e.g. Wolk *et al.*, 2013: 399), but this has the drawback that the underpopulated levels (often the majority) are considered to be the same. Though in practice, this may yield a useful model, it leads to the misrepresentation of the non-independence of the observations, flouting the very motivation of random effects.

3. Using memory-based learning (Daelemans and van den Bosch, 2005; Theijssen *et al.*, 2013; Van den Bosch and Bresnan, 2015; Pijpops, 2019; De Troij *et al.*, 2021) instead of regression analysis. The advantage is that one is no longer assuming a stable effect of higher-order notions (e.g. animacy or topicality), which, as argued by Dąbrowska (2017: 23–25, 37), are often too vague to be falsifiable and are difficult to operationalize. The drawback is that the method is relatively 'black-box', and 'does not allow an easy interpretation at a more general and abstract, linguistically meaningful, level' (Theijssen *et al.*, 2013: 249).
4. Running a separate (Distinctive) Collexeme Analysis (Stefanowitsch and Gries, 2003; Gries and Stefanowitsch, 2004), or related approaches (Schmid and Küchenhoff, 2013). The advantage is that we get a clear view on the lexical effects, but the downside is that the methods do not allow for multifactorial control (Bloem, 2021: 115), and may be prone to overfitting.

In this paper, we want to extend our methodological toolkit, and investigate the use of regularization by shrinkage methods from the field of machine learning to get a better grip on the lexical effects. More specifically, we want to use Lasso regression with cross-validation. The advantages are first that we can stick to the regression design as opposed to solution (3) and (4) above, second that we can still retain multivariate control, with extra-linguistic and intra-linguistic factors, and interactions, as opposed to solution (4) above, third that we get a generalizable method that is more robust against overfitting, as opposed to solution (1) above, and fourth that we are not stretching the use of random effects beyond what they are designed for, as opposed to solution (2) above. As an additional bonus, the method is able to increase the number of predictors that can be entered into the regression.

The potential downsides are that we have to give up on the mixed-model design for other typical random factors, like author, text, or genre (Gries, 2015). There are two pragmatic solutions: if the genre division is not too fine-grained, it may be wiser to enter it as a fixed effect, and the same goes for author and/or text, if there is only a limited number of different authors or texts attested in the sample (Speelman *et al.*, 2018: 3). Note that the binning solution alluded to above, in which all authors or texts from which only a few examples are sampled are binned in one category, is, in a sense already a step in the direction of reducing the number of factor levels. If the number of authors or texts is larger, this may become increasingly cumbersome or unfeasible, but then one might opt to avoid using multiple observations from the same text file, the second pragmatic solution. Of course, this may not always be feasible, for instance in corpus studies in which each author or text provides a fair number of observations, and the total number of observations is not abundant. This type of study is fairly typical in diachronic linguistics, if we sample from periods with a limited number of texts. But in synchronic corpus linguistics, with increasingly large corpora, often in the order of magnitude of hundreds of millions or even of billions of tokens, this may be less problematic. The typical situation, which we also have in the case study at hand (infra), is one in which some authors or texts provide many observations, and many authors or texts provide very few observations.

These pragmatic solutions are not ideal, to be sure, and it would be preferable if the regularization techniques like Lasso could be integrated in the mixed-effect approach, but at present, these techniques are cumbersome when combined with ~~n~~-fold cross-validation. Pioneering papers like Bondell *et al.* (2010), Schelldorfer *et al.* (2011), and Groll and Tutz (2014), show that advances are made to integrate random effects in penalized regularization.

2. Regularized Regression

Entering lexical lemmas as fixed effects is often not a sensible option in corpus studies for lexical effects. If the lemma is treated as a factor with all lexemes as factor levels, the model will suffer under an unwieldy proliferation of regressors. This is, in essence, a problem of 'overfitting'. Overfitting happens when a model has a tight fit to the data it was fed, but at the cost of extrapolation. The problem is not unknown in linguistics (see e.g. Hamrick, 2019), but it is often ignored in corpus studies, and when it is addressed the concern is mainly to get accurate estimates of the predictors or avoiding to fit too sensitive a high-order polynomial in a regression analysis (especially in additive mixed-models (GAMMs) that have a penalty for increasing the number of splines, see e.g. Ghyselen and Vandenberghe, 2019 for a linguistic application, and explicit reference to overfitting, p. 39), rather than pruning the predictor set itself. Still, while higher-order polynomials are only rarely used in corpus-based studies (pace recent advances in general additive modeling), variable selection is almost always an issue.

In principle, one can avoid overfitting in variable selection by using strict theory-driven variable selection. This means that you only include in the model: (a) focal predictors for which you have clear, a priori hypotheses about how and why they may affect the choice of variant, and (b) control predictors, factors that are known to affect both the outcome variable and the predictors. An example of such a control is e.g. the age of the participant when measuring the effect of reading ability (the predictor) on theory-of-mind (the outcome): younger kids will both have lower reading ability and less evolved theory-of-mind.

For many studies in corpus-based linguistics, however, such strict theory-driven variable selection is not feasible, because one does not always have clear-cut a priori hypotheses. Lexical effects are typically rather open-ended.

In machine learning the common practice is to do cross-validation by dividing the available data in a training set and a test set, and to see how well a model performs when confronted with unseen data (Ng, 2018; Deisenroth *et al.*, 2020). Cross-validation in machine learning is often done in combination with so-called regularization methods (Hastie *et al.*, 2013; Deisenroth *et al.*, 2020: 262–263), a family of techniques with Ridge regression, Lasso Regression and Elastic Net as its members. These techniques introduce a bias, also known as 'regularization'. This bias takes the form of a penalty, called 'lambda' (λ), which is multiplied with the coefficient of a (set of) predictor(s). This penalty scaled on the coefficients is then added to the regression equation which is used in the model fitting algorithm.

Why would you deliberately make the fit worse by adding a penalty? The reason is that the penalty makes the estimation of the coefficients more conservative. More conservative estimates may perform better when the model is confronted with unseen data, because lower coefficients will reduce drastic differences between the 'old/seen' and 'new/unseen' data. Unseen data are obtained by re-using the dataset we already have by applying k-fold cross-validation: the data is repartitioned k times and each time 1 – 1/k of the data is used as the training set for the model fit, and 1/k of the data as the test set. The model quality is iteratively checked against the test set. The optimal λ is established by minimizing the average deviance of the *k* test sets and their respective *k* training sets. If all *k* times, the coefficient estimation in the training set gives accurate predictions for the test set, the optimal λ can be kept low. If, on the other hand, the coefficient estimation in the training set yields a bad fit for the test set, λ will be higher. The optimal λ penalty can be so high that the coefficient is reduced to zero. This means that the variable is not helpful in predicting the outcome.

Regularization shrinkage with cross-validation is particularly useful in analyses with many potential explanatory variables. To use a non-linguistic example, suppose you want to assess what genes are responsible for hereditary differences in IQ. Ignoring the environmental effects and the gene-environment interactions, there are potentially many genes that can simultaneously have an effect on IQ. Say you have measured several hundred or thousand gene expressions in a number of individuals. Typically in such a study the number of individuals will be markedly lower than the number of genes you look at. Fitting the outcome variable IQ in a hyperplane of predictors is mathematically impossible if you have fewer observations than dimensions in the hyperplane, but cross-validation allows us to circumvent this predicament.

To turn to linguistics again, suppose we have a model where we want to predict the outcome of a binary alternating construction, e.g. the Dutch dative alternation, by length of the recipient and region (Belgium [0] vs. the Netherlands [1]) (to be sure, this is an oversimplified model for expository purposes). A straightforward research design will be to add a random intercept for ‘verb’ (see above) as we do not want to assume that *geven* (‘give’), behaves in exactly the same way as *vertellen* (‘tell’), *sturen* (‘send’), *overhandigen* (‘hand’), etc. This would amount to a model in (1) (fitted through a maximum likelihood estimation), where $x_{i,j}$ stands for the j^{th} observation of verb i, β_0 is the model intercept, β_1 the weight for the recipient length (*RecLength*), β_2 the weight if the observation is from the Netherlands (*ND*), and v_i is the by-verb correction (for simplicity’s sake, we will not discuss random slopes here). The model will tell us the effect of pronominality and region, correcting for the accidental set of verbs that we have in our dataset. Such a model, while decidedly better than a model with fixed effects only, as in (2), does not tell us, however, whether some verbs are more relevant for the model than others.

To do this, we can turn to Lasso. We need to slightly transform our data, as will be illustrated below, and siphon the verbs over to the fixed-effect structure. Each verb is now entered as a categorical binary predictor. Lasso will use a penalty λ on the absolute value of the coefficient of each of the predictors, see (3), in which n stands for the number of different verbs, i identifies the verb and $v_{i,j}$ the value of the pseudo-observation j of verb i. The advantage of Lasso regression over Ridge regression is that it can shrink coefficient(s) all the way to zero, under an optimal λ. Coefficients shrunken to zero are not retained by the model, effectively carrying out a variable selection.

(1) $\ln(odds(x_{i,j} = PrepDat)) = \beta_0 + \beta_1 RecLength_{i,j} + \beta_2 ND_{i,j} + v_i \; (v_i \sim N(0,\sigma_i^2))$

(2) $\ln(odds(x_j = PrepDat)) = \beta_0 + \beta_1 RecLength_j + \beta_2 ND_j$

(3) $\ln(odds(x_j = PrepDat)) = \beta_0 + \beta_1 RecLength_j + \beta_2 ND_j + \Sigma_{i=1}^{n} \beta_{3,i} v_{i,j} +$
$\lambda(|\beta_1| + |\beta_2| + \Sigma_{i=1}^{n} |\beta_{3,i}|)$

3. A Real-life Example

To illustrate how Lasso regression works in real life, we used an existing dataset on the Dutch alternation in the verb *zoeken*, which can be realized with a direct object as in (4) or with a prepositional object, as in (5) (not unlike English *search (for)*), see Haeseryn *et al.* (1997: 1168).

(4) *Zoek je je paraplu?*
search you your umbrella
'Are you looking for your umbrella?'

(5) *Zoek je naar je paraplu?*
search you to your umbrella
'Are you looking for your umbrella?'

Suppose you want to investigate whether the alternation is lexically entrenched, and depends on the head noun of the theme. Maybe *paraplu* ('umbrella') prefers the prepositional construction, but another noun, say *kat* ('cat'), prefers the transitive construction. However, you may lack clear *a priori* expectations about how or why certain nouns would be entrenched in the alternating variants. That is, you may not have a hypothesis that predicts exactly which objects will prefer the prepositional construction, and which prefer the transitive construction. This is exactly the kind of lexical effect that is central to construction grammar (see Pijpops, 2019 for an in-depth study), but it will serve for our expository purposes here.

To investigate this, we took an existing dataset (Pijpops, 2019), based on the 500 million token Open SoNaR corpus (Oostdijk *et al.* 2013). To avoid the issue of integrating a random factor for Text File, we took one observation of each text, as most texts yielded a limited number of observations anyway: in the original dataset 78% of the texts (45,255/58,065) only yielded one hit, and 91% (53,011/58,065) of the texts yielded five observations or less. Amalgamating these 53,011 texts in one random factor is not ideal, so we went for the second 'pragmatic solution' mentioned above. Furthermore, we ignored pronominally realized Themes, and only retained observations with Theme lemmas that occur at least 10 times. This yielded a dataset comprising 33,528 observations of the verb *zoeken*, of which 27,915 sport the transitive variant, and 5,613 the prepositional variant. This binary factor (Variant) is the outcome variable in our regression analysis.

For each observation, we also have information about the complexity of the Theme argument (Theme Complexity), calculated as the natural logarithm of its number of words, as well as the position of the Theme argument (before or after the verb) (Theme position), and country (Belgium vs. the

Netherlands), and the head lemma of the Theme argument (Theme Lemma). We know from earlier research (Pijpops *et al.*, 2018; Pijpops, 2019) that the choice is partially dependent on the Theme lemma, so we want to look into its lexical-semantic effects, controlling for Theme Complexity and Theme Position, and Country. In order to do this, we transformed the dataset so that Theme Lemma is no longer one factor with 554 levels, but 554 different binary factors: either the specific Theme Lemma occurs in a particular observation, or it does not. We now have a dataset of 33,258 rows (all different texts as rows) and 558 columns: 554 Theme Lemmas, plus a value for Theme Complexity, Theme Position and Country. This dataset comprises far too many variables to run an adequate binomial regression model on the constructional variant. By a common rule of thumb, the maximal number of regressors is 1/20 of the number of observations of the least frequent outcome level. In our case, we have double the number of regressors. This is where regularization with cross-validation, in the form of Lasso regression, offers a solution.

For the analysis, we will use the R package glmnet (Friedman *et al.*, 2010).[1] Running a 10-fold cross-validated Lasso regression in which we binomially regress the Variant (direct object vs. prepositional object) on the different theme lemma levels (now entered as binary predictor variables), controlling for the Theme Complexity, Theme Position and Country covariates, we arrive at an optimal λ of 0.0007.

Of the 554 Theme lemmas, the coefficients of 73 are reduced to zero. Of the other covariates, Theme Complexity, Theme Position and Country are also retained as significant predictors (see Appendix A).[2]

Let us focus on the lexical effects. A concern might be that the Lasso regularization shrinks Theme lemmas purely on the basis of their frequencies, but this does not appear to be the case: a binomial regression with a binary outcome variable Retained by the Lasso, and the Attested Frequency of the Theme lemma (i.e. the number of times the Theme Lemma occurs in the dataset) as the predictor, does not reach significance (p = 0.626).

We now have a list of 481 Theme lemmas retained by the Lasso, which can be subjected to further analysis. First off, we compare the regularized Lasso coefficient, obtained under multivariate control, with the results of a Distinctive Collexeme Analysis we carried out on the same dataset (using the collostructions package in R, Flach, 2021). The collexeme attraction strength can be assessed with different association measures. We will use the default Log Likelihood measure, which is sensitive to frequency and uses significance values, and the Odds Ratio (OR), which is frequency-insensitive and uses effect size. The Odds Ratios theoretically range from −infinity to +infinity and can be directly compared to the coefficients in the Lasso Regression. For the Log-Likelihood, we reversed the sign of the collostructional strength when the attraction was towards the direct object variant. A large negative value thus signifies a strong attraction towards the direct object construction, a large positive value signifies a strong attraction towards the prepositional object construction, and a value close to zero means the Theme Lemma is not particularly attracted to either variant.

The Lasso coefficients show a near-perfect correlation with the OR-based collostructional strength (Pearson correlation = 0.98, p < 0.001). For the Log-Likelihood-based collostructional strength, the correlation is weaker, but still sizeable (Pearson correlation = 0.52, p < 0.001), see also Figure 1. Table 1 shows the classification agreement (which is the same for both types of Distinctive Collexeme Analysis). Ignoring the Theme Lemmas not retained by the Lasso regression (the last row in Table 1), the agreement in the classification is 95.2%.

Set off against a well-known technique in corpus linguistics, the Lasso regression seems to behave as expected. We take this as an indication of the quality of our method. In order to further test the differential merits of the three approaches, we looked deeper into the lexical effects.

It is hard to pin down exactly what determines the choice for either constructional variant at the lexical level. One reasonable assumption is that Theme Lemmas that are semantically close to one another have a predilection for the same Variant. That is, synonyms and near-synonyms, such as *contact* ('contact') and *aansluiting* ('connection'), would prefer the same argument structure (Pijpops *et al.*, 2018: 535). Let us call this the 'birds of a feather flock

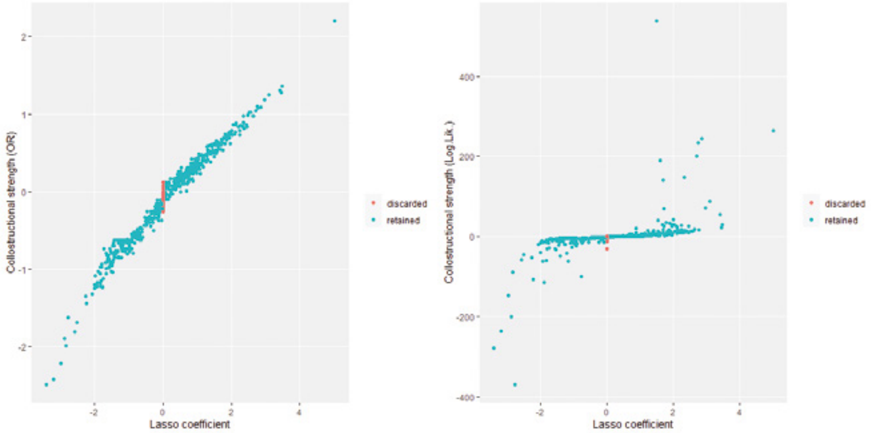

Figure 1: Correlation between the Lasso coefficient and the OR-based collostructional strength (OR) (left) and the Lasso coefficient and the Log-Likelihood-based collostructional strength (right). Each dot represents a different Theme Lemma. Color coding for whether or not the Theme Lemma is retained by the Lasso.

Table 1: Agreement in constructional preference: Lasso vs. Distinctive Collexeme Analysis

Direct Object		**Constructional preference (Distinctive Collexeme Analysis)**	
		Direct Object	**Prep. Object**
Constructional preference (Lasso)	Direct Object	281	0
	Prep. Object	23	240
	Discarded (shrunken to zero)	71	2

together' effect. It is certainly not a strict law, but merely a (possibly hard to eyeball) tendency, as near-synonyms can be attracted to different variants (see also Diessel, 2019, Ch.7 and Gries and Stefanowitsch, 2004). Can we operationalize this effect in a statistically more informed way?

We have taken the semantic vectors for all Theme lemmas that were retained by the Lasso as represented in the Snaut repository (http://meshugga.ugent.be/snaut-dutch/). These vectors are constructed on the 500 million SoNaR corpus and a corpus of subtitles (Mandera *et al.*, 2017). Not all Theme Lemmas are represented in the repository, but 95% (458 out of the 481) are. We used these vectors to build a matrix of cosine distances. This matrix was then turned into a dendogram, using the Ward clustering method. Then we made 457 cluster groupings with increasing granularity: from a macro-cluster with two groups to a micro-cluster with 458 groups. In this last group, all Theme Lemmas are in their own cluster. For each of these clusters, we then ran three regression analyses: one regressing the Lasso coefficient on the cluster membership, another regressing the OR-based collostructional strength on the cluster membership, and the last one regression the Log-Likelihood-based collostructional strength on the cluster membership. The rationale behind the regression analyses is that the choice for an object is partially determined by group membership in the cluster. For all 1,371 (457 * 3) regression analyses, we extracted the R^2 value. Obviously, the last 458-groups cluster will have a perfect R^2 for both regression analyses.

How do the two methods fare? In Figure 2 the x-axis gives the number of groups in the hierarchical cluster, and the y-axis gives the R^2 value for the three methods. See Levshina and Heylen (2014) and Pijpops (2019) for a related approach. As can be appreciated, the Lasso methods progresses in lock-step with the OR-based Distinctive Collexeme Analysis, and both consistently outperform the Log-Likelihood-based Distinctive Collexeme Analysis. Apart from showing that the Lasso regression performs adequately, we take this as support that effect-size-based Collostructional methods are preferable over significance-based Collostructional methods.

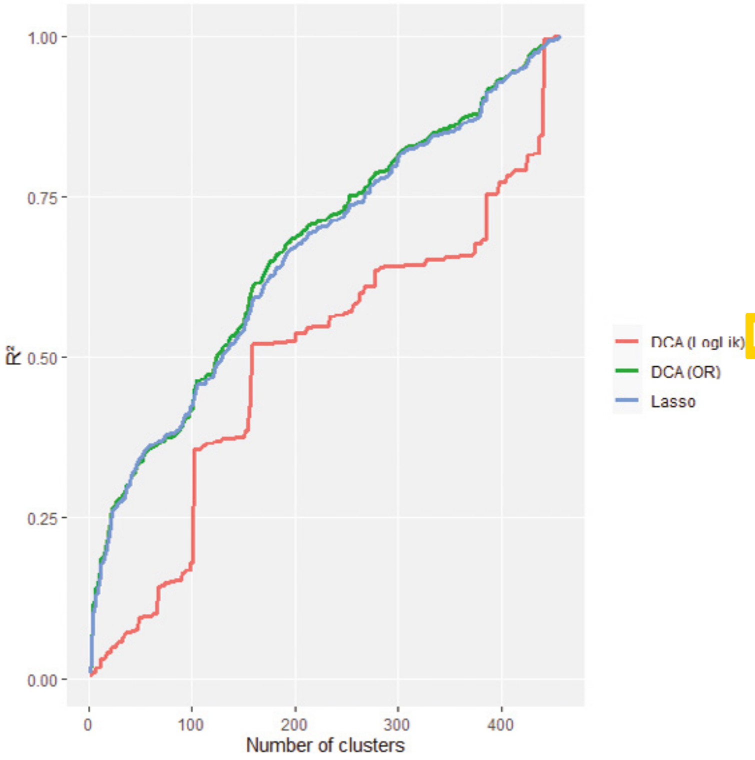

Figure 2: R^2 values of linear regression models, predicting the Lasso coefficient and the collostructional strength of an OR-based Distinctive Collexeme Analysis and a Log-Likelihood-based Distinctive Collexe Analysis,, all on the y-axis, on cluster membership for clusters of increasing granularity (x-axis)

4. Conclusions

Lasso regression, a regularization technique from the field of machine learning, can be used for assessing lexical effects in syntax. Its advantage over Collostructional analysis techniques is that it works under multivariate control. Like Collostructional analysis, it is computationally relatively light: all analyses in this paper have been carried out on a conventional laptop with open-source software.

One potential downside of the method as introduced in this short paper is that it does not take into account random effects. This is not an insurmountable problem. Lasso regression is currently extended to mixed models as well (e.g. Groll and Tutz, 2014), and users of the R software may fruitfully apply the glmmLasso package for R (Groll, 2017). We leave this extension for a future paper. The mathematics are more complex, the method is computationally much heavier, and finding the optimal lambda penalty is not as straightforward. The coming years are likely to see advances in dealing with overfitting in mixed-models. A pioneering paper is Roberts *et al.* (2017), but the issue is difficult to accommodate.

We think the fixed-effect Lasso regression we employed in this paper strikes a reasonable balance between complexity and useability, especially because the technique relaxes the restrictions on the number of regressors the regression can handle. The approach advocated in the present project is not a complete overhaul of the field by discarding well-established methods, but rather by enriching them with machine learning tools (see also Yarkoni and Westfall, 2017).

Acknowledgements

We would like to thank Benedikt Heller, Dirk Speelman, Stefano De Pascale, and the two anonymous reviewers for insightful comments on earlier versions of this article.

About the Authors

Freek Van de Velde (KU Leuven) is associate professor of Dutch linguistics and historical linguistics. His research focuses on quantitative approaches to variation and change and evolutionary linguistics. He received his PhD in 2009, with a work on the diachrony of the noun phrase.

Dirk Pijpops (University of Liège) works as lecturer of Dutch. He is affiliated with the research unit Lilith. His research focuses language variation and change, which he studies in order to answer questions in usage-based theoretical linguistics. Methodologically, his work builds on quantitative corpus analyses and agent-based computer simulations. He received his PhD in 2019 at the University of Leuven, with a thesis focused on argument structure variation in Dutch.

Notes

1. For data wrangling, we used the tidyverse tools (Hadley *et al.*, 2019).
2. At first sight, it may come as a surprise that Theme Complexity has a negative effect on the use of the prepositional variant. One would expect the longer variant to be used in more complex environments. The reason is the presence of the covariate for Theme Position. As argued in Pijpops *et al.* (2018), complex Theme arguments indeed eschew the preposition when the argument in placed before the verb, as a result of information distribution considerations.

References

Bloem, Jelke (2021). *Processing verb clusters*. Utrecht: LOT Dissertation Series.

Bondell, Howard D., Arun Krishna, and Sujit K. Ghosh (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics* 66(4): 1069–1077. https://doi.org/10.1111/j.1541-0420.2010.01391.x

Bresnan, Joan, Anna Cueni, Tatiana, and R. Harald Baayen (2007). Predicting the dative alternation. In Gerlof Bouma, Irene Kraemer, and Joost Zwarts (Eds), *Cognitive Foundations of Interpretation*. Amsterdam: Royal Netherlands Academy of Science. 69–94.

Bresnan, Joan and Ford, Marilyn. (2010). Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86: 168–213. https://doi.org/10.1353/lan.0.0189

Cappelle, Bert (2006). Particle placement and the case for 'allostructions'. In Doris Schönefeld (Ed.), *Constructions all Over: Case Studies and Theoretical Implications*. [Special issue of *Constructions*].

Colleman, Timothy (2006). *De Nederlandse datiefalternantie. Een constructioneel en corpusgebaseerd onderzoek*. PhD Dissertation. UGent.

Dąbrowska, Ewa (2017). *Ten Lectures on Grammar in the Mind*. Leiden: Brill. https://doi.org/10.1163/9789004336827

Daelemans, Walter and Antal van den Bosch (2005). *Memory-based Language Processing*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511486579

Deisenroth, Marc P., A. Aldo Faisal, and Cheng Soon Ong (2020). *Mathematics for Machine Learning*. Preprint book. https://mml-book.github.io/ https://doi.org/10.1017/9781108679930

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 Belgian and Netherlandic Dutch. *Natural Language Engineering*. https://doi.org/10.1017/S1351324921000097

De Vaere, Hilde (2020). *The ditransitive alternation in present-day German. A corpus-based analysis*. PhD Dissertation. UGent.

Diessel, Holger (2019). *The Grammar Network: How Linguistic Structure is Shaped by Language Use*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108671040

Flach, Susanne (2021). *Collostructions: An R Implementation for the Family of Collostructional Methods*. R package version 0.2.0.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33(1): 1–22. https://doi.org/10.18637/jss.v033.i01

Ghyselen, Anne-Sophie, and Roxane Vandenberghe (2019). Over *etwat*, *etwuk* en *iets*:geografie en dynamiek van het onbepaald voornaamwoord voor zaak in West-Vlaanderen. *Taal en Tongval* 71(1): 31–60. https://doi.org/10.5117/TET2019.1.GHYS

Goldberg, Adèle (2006). *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.

Gries, Stefan Th. (2000). *Towards multifactorial analyses of syntactic variation: the case of particle placement*. PhD Dissertation, University of Hamburg.

Gries, Stefan Th. and Anatol Stefanowitsch (2004). Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9(1): 97–129. https://doi.org/10.1075/ijcl.9.1.06gri

Gries, Stefan Th. (2015). The most underused statistical method in corpus linguistics: multi-level (and mixed-effects) models. *Corpora* 10(1): 95–125. https://doi.org/10.3366/cor.2015.0068

Groll, Andreas (2017). *glmmLasso: Variable Selection for Generalized Linear Mixed Models by L1-Penalized Estimation. R package version 1.5.1*. https://CRAN.R-project.org/package=glmmLasso.

Groll, Andreas and Gerhard Tutz (2014). Variable selection for generalized linear mixed models by L1-penalized estimation. *Statistics and Computing* 24(2): 137–154. https://doi.org/10.1007/s11222-012-9359-z

Grondelaers, Stefan (2000). *De distributie van niet-anaforisch er buiten de eerste zinplaats: sociolexicologische, functionele en psycholinguïstische aspecten van* er*'s status als presentatief signaal*. PhD Dissertation, KU Leuven.

Pijpops, Dirk (2019). *Where, how and why does argument structure vary? A usage-based investigation into the Dutch transitive-prepositional alternation*. PhD Diss. KU Leuven.

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. https://doi.org/10.1017/langcog.2018.13

Haeseryn, Walter, Kirsten Romijn, Guido Geerts, Jaap de Rooij, and Maarten van den Toorn (1997). *Algemene Nederlandse Spraakkunst*. 2nd end. Groningen: Nijhoff.

Hamrick, Phillip (2019). Adjusting regression models for overfitting in second language research. *Journal of Research Design and Statistics in Linguistics and Communication Science* 5(1-2): 107–122. https://doi.org/10.1558/jrds.38374

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2013). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. 2nd edn. Berlin: Springer.

Klavan, Jane and Dagmar Divjak (2016). The cognitive plausibility of statistical classification models: Comparing textual and behavioral evidence. *Folia Linguistica* 50: 355–384. https://doi.org/10.1515/flin-2016-0014

Levshina, Natalia and Kris Heylen (2014). A radically data-driven construction grammar: experiments with Dutch causative constructions. In Ronny Boogaart, Timothy Colleman, and Gijsbert Rutten (Eds), *Extending the Scope of Construction Grammar*. Berlin: Mouton de Gruyter. 17–46. https://doi.org/10.1515/9783110366273.17

Mandera, Paweł, Emmanuel Keuleers, and Marc Brysbaert (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: a review and empirical validation. *Journal of Memory and Language* 92: 57–78. https://doi.org/10.1016/j.jml.2016.04.001

Ng, Andrew (2018). *Machine learning yearning*. E-book. https://d2wvfoqc9gyqzf.cloudfront.net/content/uploads/2018/09/Ng-MLY01-13.pdf

Oostdijk, Nelleke, Martin Reynaert, Véronique Hoste, and Ineke Schuurman (2013). The construction of a 500 million word reference corpus of contemporary written Dutch. In Peter Spyns and Jan Odijk (Eds), *Essential Speech and Language Technology for Dutch: Results by the STEVIN-programme*, 219–247. Berlin/Heidelberg: Springer. https://doi.org/10.1007/978-3-642-30910-6_13

Perek, Florent (2015). *Argument Structure in Usage-based Construction Grammar*. Amsterdam: John Benjamins. https://doi.org/10.1075/cal.17

Rappaport-Hovav, Malka and Beth Levin (2008). The English dative alternation: The case for verb sensitivity, *Journal of Linguistics* 44: 129–167. https://doi.org/10.1017/S0022226707004975

Roberts, David R. Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* 40: 913–929. https://doi.org/10.1111/ecog.02881

Röthlisberger, Melanie (2018). *Regional variation in probabilistic grammars: a multifactorial study of the English dative alternation*. PhD Dissertation. KU Leuven.

Röthlisberger, Melanie, Jason Grafmiller, and Benedikt Szmrecsanyi (2017). Cognitive indigenization effects in the English dative alternation. *Cognitive Linguistics* 28(4): 673–710. https://doi.org/10.1515/cog-2016-0051

Schelldorfer, Jürg, Peter Bühlmann, and Sara van de Geer (2011). Estimation for high-dimensional linear mixed-effects models using L1-Penalization. *Scandinavian Journal of Statistics* 38: 197–214. https://doi.org/10.1111/j.1467-9469.2011.00740.x

Schmid, Hans-Jörg and Helmut Küchenhoff (2013). Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and cognitive underpinnings. *Cognitive Linguistics* 24(3): 531–577. https://doi.org/10.1515/cog-2013-0018

Speelman, Dirk (2014). Logistic regression: A confirmatory technique for comparisons in corpus Linguistics. In Dylan Glynn and Justyna A. Robinson (Eds), *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*. 487–533. Amsterdam: John Benjamins. https://doi.org/10.1075/hcp.43.18spe

Speelman, Dirk, Kris Heylen, and Dirk Geeraerts (2018). 'Introduction'. In: Dirk Speelman, Kris Heylen and Dirk Geeraerts (Eds), *Mixed-effects Regression Models in Linguistics*. 1–10. Cham: Springer. https://doi.org/10.1007/978-3-319-69830-4_1

Stefanowitsch, Anatol and Stefan Th. Gries (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2): 209–244. https://doi.org/10.1075/ijcl.8.2.03ste

Theijssen, Daphne, Louis ten Bosch, Lou Boves, Bert Cranen, and Hans van Halteren (2013). Choosing alternatives: Using Bayesian networks and memory-based learning to study the dative alternation. *Corpus Linguistics and Linguistic Theory* 9: 227–262. https://doi.org/10.1515/cllt-2013-0007

Van den Bosch, Antal and Joan Bresnan (2015). Modeling dative alternations of individual children. *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning*.103–112. https://doi.org/10.18653/v1/W15-2414

Van de Velde, Freek, Stefano De Pascale, and Dirk Speelman (Forthcoming). Generalizability in mixed models: Lessons from corpus linguistics (response article). *Behavioral and Brain Sciences*.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani (2019). Welcome to the tidyverse. *Journal of Open Source Software* 4(43): 1686. https://doi.org/10.21105/joss.01686

Winter, Bodo (2020). *Statistics for Linguistics. An Introduction Using R*. New York: Routledge.

Wolk, Christoph, Joan Bresnan, Anette Rosenbach, and Benedikt Szmrecsanyi (2013). Dative and genitive variability in Late Modern English: exploring cross-constructional variation and change. *Diachronica* 30(3): 382–419. https://doi.org/10.1075/dia.30.3.04wol

Yarkoni, Tal and Jacob Westfall (2017). Choosing prediction over explanation in psychology: lessons from machine learning. *Perspectives on Psychological Science* 12(6): 1100–1122. https://doi.org/10.1177/1745691617693393

Zehentner, Eva (2019). *Competition in Language Change: The rise of the English Dative Alternation*. Berlin: De Gruyter. https://doi.org/10.1515/9783110633856

Appendix

Lasso coefficient estimates (on the logit scale). Positive estimates signify a pull towards the prepositional object Variant, negative estimates signify a pull towards the direct object Variant. A coefficient of zero means that the variable has no discriminatory effect.

Intercept	–3.52
Theme Complexity	–0.07
Theme Position	0.86
Country (The Netherlands)	0.21
aandacht	–1.31
Aandeel	1.86

aandeelhouder	–0.35
aanknopingspunt	1.04
aanleiding	0.40
aannemer	–0.28
Aanpak	2.04
aansluiting	–2.23
Aantal	–0.67
Aanval	–1.83
aanvaller	1.33
aanwijzing	2.97
Accoord	0.34
achterpoortje	0.00
achtergrond	1.60
Acteur	–0.25
Activiteit	0.61
Actrice	0.00
Adem	2.09
adoptieouders	–1.24
Adres	0.00
Advies	–1.15
advocaat	–0.62
afkoeling	–1.84
Afleiding	–0.75
Afnemer	–0.29
afwisseling	–1.39
afzetmarkt	–0.41
Agent	–1.59
Alibi	0.87
alternatief	1.70
Ander	0.93
antwoord	0.49
appartement	–0.64
arbeidskracht	0.71
Arbeider	–0.31

argument	1.56
Artiest	0.09
Asiel	–0.97
Assistent	0.00
Auto	1.29
avontuur	–0.87
Baan	0.00
Baantje	–0.75
Baas	0.00
Baasje	–1.58
Bal	0.28
Balans	0.89
Band	0.92
Basis	–0.97
Bedrijf	0.00
Beeld	0.69
Beetje	0.00
begeleider	–1.45
begeleiding	–0.12
Begrip	0.44
behandeling	1.51
Belang	0.32
Belg	0.07
benadering	0.81
bescherming	–2.58
beschutting	–1.27
besparing	1.14
Bestaan	–1.38
bestemming	0.09
bestuurder	0.52
betekenis	1.27
betrekking	0.17
bevestiging	0.00
Bewijs	1.67

Bewoner	0.54
bezigheid	0.00
Bijstand	–1.10
bijverdienste	–1.19
Bloed	–0.13
Bodem	0.00
Boek	0.00
Boel	–0.93
bondgenoot	0.39
Broek	0.72
Broer	1.92
Bron	0.60
Bruid	–1.04
Buit	1.11
Cadeau	–1.34
cadeautje	0.00
Cafe	–0.19
Cd	–0.06
chauffeur	–0.48
Club	–0.92
Coach	0.00
Coalitie	1.15
coalitiepartner	–0.30
combinatie	1.06
compensatie	0.74
compromis	1.40
computer	0.87
Concept	2.53
Conflict	–1.66
confrontatie	–1.54
consensus	1.18
Contact	–2.78
Contrast	0.88
controverse	–1.19

Dader	0.85
Dame	–0.08
Datum	1.08
Deel	–0.19
deelnemer	–0.82
Dekking	–1.57
Detail	2.57
Dialoog	–1.66
Diamant	1.96
Dief	0.60
diepgang	–0.16
Diepte	0.26
Dier	0.40
Ding	1.04
directeur	–0.86
Docent	–1.73
Dochter	0.67
document	1.83
Doel	0.00
Doelman	0.00
doelpunt	1.27
Donor	0.38
Dood	–1.51
draagvlak	0.45
drenkeling	3.47
Eenheid	0.98
eensgezindheid	0.73
eerherstel	0.00
Eet	1.10
Effect	0.05
Eigenaar	–1.32
Element	1.29
Emplooi	–1.44
erkenning	0.00

Euro	0.00
evenwicht	0.70
Excuus	0.00
exemplaar	0.00
Expansie	–0.81
Explosief	3.45
fabrikant	1.06
Familie	0.00
Feit	1.28
Fiets	0.68
Figurant	–2.07
financier	–0.13
financiering	0.00
Flat	–1.17
Fonds	–0.50
Formule	2.03
Foto	0.16
fotograaf	–1.27
Fout	–0.09
Frank	–0.03
Functie	0.23
gastgezin	–1.91
Gat	0.63
Gaatje	2.21
geborgenheid	0.00
Gebouw	0.00
Gegeven	–0.23
Geheim	0.18
Geld	0.21
geldschieter	0.00
gelegenheid	0.93
Gelijk	–0.18
gelijkenis	–0.36
gelijkmaker	1.39

Geluid	2.13
Geluk	–1.84
gerechtigheid	–0.18
Getuige	–1.05
gezelschap	–1.99
Gezicht	1.05
Gezin	–0.89
Gids	0.16
God	0.04
Goed	1.56
Goud	2.24
Graf	1.04
Grens	0.78
Groei	0.00
Groep	–0.26
Grond	0.05
Hand	–1.67
handtekening	0.36
harmonie	1.77
heenkomen	–1.75
Heil	–3.41
Hobby	–0.64
Hond	0.79
honderden	–1.31
Hotel	0.36
Houding	1.14
Houvast	–0.04
Huis	–0.56
huisvesting	–0.13
Hulp	–2.89
Humor	0.00
huurwoning	0.75
Huurder	0.00
Idee	1.54

identiteit	2.01
Iemand	–1.18
Imago	–1.29
Info	–1.46
informatie	0.28
ingenieur	–0.03
ingrediënt	1.88
Inkomst	0.43
inspiratie	–2.84
invalhoek	1.01
investeerder	0.41
investering	–1.54
Invulling	1.85
Inwoner	–1.42
Job	0.00
jobstudent	–1.97
Jongen	1.40
Jongetje	1.13
Jongere	–0.50
Kamer	0.00
Kanaal	–0.34
kandidaat	–0.14
Kans	1.33
Kapitaal	0.09
Keeper	0.26
Kern	0.93
Kick	–1.48
Kind	0.24
kinderopvang	–0.84
Klant	–1.29
Kleding	2.26
Kleed	0.64
kompaan	0.52
Koper	–0.02

Koppel	–0.26
Kracht	–1.38
kunstenaar	–0.85
Kwaliteit	0.10
Land	0.00
Leider	0.36
Leven	0.57
leverancier	0.55
Lichaam	3.41
Lid	–0.53
Lief	–1.05
Liefde	–0.32
Lijk	2.65
Lijn	1.49
Link	0.00
Locatie	0.76
Logica	–0.28
Logies	0.00
Logo	–0.15
lokaal	–0.05
lokatie	0.00
maatregel	2.24
man	0.00
manager	–0.90
manier	1.60
markt	–0.38
materiaal	1.15
medestander	1.17
medewerker	–0.91
medicijn	1.95
medium	0.00
meerderheid	0.52
meerwaarde	0.16
meisje	1.24

mens	–0.23
meter	0.00
methode	1.98
middel	1.54
midden	–1.14
middenweg	0.26
miljard	–0.66
miljoen	–0.29
minister	0.00
mix	1.80
Mladic	–0.92
model	–0.73
moeder	0.44
moeilijkheid	–1.57
mogelijkheid	2.75
moment	1.71
monitor	–1.83
moordenaar	–0.15
motief	0.72
motivatie	–0.88
muziek	0.65
muzikant	–1.78
naam	0.74
niche	–0.39
nieuws	0.99
noodoplossing	1.54
nuance	–1.52
nummer	0.32
object	–0.12
olie	2.37
onderdak	–1.93
onderdeel	1.31
onderkomen	–1.22
onderneming	1.01

onderwerp	0.67
ontspanning	–0.40
ontwerper	–1.53
oogcontact	–1.93
oorsprong	0.62
oorzaak	0.88
openbaarheid	–1.04
opening	1.40
oplossing	1.50
opname	–0.36
opportuniteiten	1.43
opvang	–0.37
opvanggezin	–1.32
opvangplaats	0.99
opvolger	0.48
ouder	0.00
overeenkomst	2.33
overleef	5.01
overname	1.24
overnemer	0.59
overnemers	1.00
paard	0.88
paasei	–0.98
pad	0.84
pand	0.85
parallel	2.10
parkeerplaats	0.75
parking	1.81
partij	0.60
partner	–0.09
passagier	0.83
patroon	1.36
personeel	–0.26
personeellid	–1.64

persoon	0.27
peters	–1.47
plaats	0.14
plaatsje	–0.80
plant	1.02
pleeggezin	–0.46
plek	0.00
plekje	–0.99
ploeg	–1.11
positie	0.00
presentator	–0.09
prijs	1.02
privéinvesteerder	–1.46
privépartner	0.00
probleem	–0.73
producent	0.51
product	1.72
profiel	1.37
programma	0.68
project	1.54
prooi	0.69
publiciteit	–2.01
publiek	–1.78
punt	0.68
raad	–1.27
recept	0.49
recht	–1.33
rechtsachter	0.91
redding	0.08
reden	0.68
regeling	1.75
regisseur	0.00
relatie	0.20
remedie	2.43

rendement	0.50
respect	–0.69
rest	2.47
restaurant	–0.12
revanche	–0.18
richting	–1.54
risico	–1.21
ritme	1.88
roem	–1.19
rol	0.00
route	0.32
ruimte	–0.14
rust	–0.75
ruzie	–1.81
samenhang	1.44
samenwerking	0.00
samenwerkingsverband	1.32
schaduw	0.21
schat	0.94
schilderij	1.32
schildpad	–0.69
schoen	0.00
school	–1.44
schoonheid	0.54
schuilplaats	–1.66
schuld	–1.46
schuldig	0.79
seks	0.47
sensatie	0.62
sfeer	1.19
site	0.34
situatie	1.58
slaapplaats	–0.60
slachtoffer	1.73

sleutel	0.57
sluipweg	0.00
soelaas	–2.00
soort	–0.13
spanning	0.00
speld	0.96
speler	–0.48
spijker	–1.76
spits	–0.09
sponsor	–0.18
sponsoring	0.00
spoor	2.86
spul	0.12
stabiliteit	0.94
stad	0.82
steen	0.88
stek	0.00
stem	0.10
steun	–1.89
stijl	1.10
stilte	0.42
stoel	0.00
stok	0.00
strategie	0.92
structuur	1.93
student	–0.23
stuk	1.58
stukje	0.00
succes	–0.01
systeem	2.12
taal	0.48
talent	–0.20
techniek	2.32
tegenstander	–0.62

teken	2.36
tekst	0.00
terrein	0.00
thema	1.10
thuis	–0.77
ticket	0.90
tip	–1.37
titel	0.88
toegang	–0.10
toekomst	0.00
toenadering	–2.98
toepassing	1.82
toevlucht	–3.20
topman	1.17
trainer	–0.92
trend	1.37
troost	–2.26
type	0.45
uitweg	0.19
uitbater	–0.10
uitbreiding	0.61
uitdaging	–0.04
uitgang	0.00
uitgever	–0.25
uitlaatklep	–1.61
uitleg	1.68
uitvlucht	–0.68
vader	0.00
vakman	0.00
veiligheid	–0.06
vent	–1.53
verantwoordelijk	0.00
verantwoordelijkheid	–1.05
verband	1.13

verbetering	1.97
verbinding	0.96
verdachte	1.89
verdediger	–0.14
vereniging	–1.22
verfrissing	–1.69
vergelijk	0.41
vergelijking	0.50
verhaal	0.40
verklaring	0.87
verkoeling	–2.52
verkoper	–1.42
vermist	3.10
vernieuwing	1.92
verontschuldiging	0.94
verschil	–0.47
versterking	0.67
vertaling	2.78
vertier	–1.77
vertrouwen	–1.03
vervangster	–0.97
vervanger	0.00
vervanging	2.63
vestiging	–1.23
vestigingplaats	1.10
vijand	0.00
vingerafdruk	2.46
visie	0.76
voedsel	0.70
voetbalgeluk	–1.09
volk	–0.46
voorbeeld	1.63
voordeel	–0.14
voorwerp	0.85

voorzitter	–0.48
vorm	2.71
vrede	0.00
vriend	–0.42
vriendin	0.00
vrijheid	–1.43
vrijwilliger	–1.41
vrouw	–0.18
vrouwtje	–1.10
waarde	1.26
waarheid	1.79
wagen	0.00
wapen	1.95
warmte	–1.61
water	0.92
weg	–0.76
werk	0.00
werkkracht	0.11
werkgever	1.05
werknemer	–1.07
winkel	–1.22
winnaar	–0.13
winst	0.00
woning	0.00
woonruimte	0.00
woonst	0.17
woord	2.34
wortel	0.77
wraak	0.00
zaak	1.43
zaal	–0.14
zanger	–0.11
zangeres	–1.19
zekerheid	–0.03

zender	1.02
zesletterwoord	–1.52
ziel	0.00
zin	1.43
zingeving	1.97
zon	0.00
zondebok	–0.03
zoon	0.00