Papers of the Linguistic Society of Belgium

17 | 2023

Distributional semantics and the English dative alternation: recipient and theme slots matter (to some extent)

Chiara Paolini, Benedikt Szmrecsanyi, Mariana Montes & Hubert Cuyckens

KU Leuven

URL: https://sites.uclouvain.be/bkl-cbl/en/journals/papers-of-the-lsb/volume-17-2023/paolini-szmrecsanyi-montes-cuyckens/

Electronic reference:

Paolini, Chiara, Szmrecsanyi, Benedikt, Montes, Mariana & Cuyckens, Hubert. (2023). Distributional semantics and the English dative alternation: recipient and theme slots matter (to some extent). In Marie Steffens & Thomas Hoelbeek (eds.), *Papers of the Linguistic Society of Belgium* [online] 17, 16-32. DOI: https://doi.org/10.61430/CSGM7397

Distributional semantics and the English dative alternation: recipient and theme slots matter (to some extent)

Chiara Paolini, Benedikt Szmrecsanyi, Mariana Montes, Hubert Cuyckens

KU Leuven

The dative alternation in English is one of the most extensively investigated alternations. Many quantitative studies to date have focused on more traditional formal predictors (e.g., complexity, constituent weight) to explain the choice of one variant over the other. In contrast, semantic predictors have been given relatively short shrift in variationist alternation research due to their perceived cost inefficiency. The objectives of this research are on the one hand to determine the importance of fine-grained semantic properties of the theme and recipient nouns for predicting variant choice, and to check on the other hand whether they add to the explanatory power of traditional formal predictors. To accomplish our aims, we make use of automatically generated semantic predictors using distributional models of meaning (Lenci 2018). Analysis shows that while distributional semantics predictors have significant predictive power, traditional predictors are subtly more powerful. Nevertheless, interactions between semantic and formal predictors clearly emerge from statistical analyses, opening the research to further applications and developments.

1. Introduction

This paper is concerned with the English dative alternation, one of the most well-investigated alternation in various subfields of linguistics. To encode dative relations, speakers and writers of English may use two structural patterns, the ditransitive dative, as in (1a), or the prepositional dative, as in (1b), which according to many analysts are semantically/functionally broadly equivalent (see Bresnan et al. 2007 for discussion).

(1) a. *The ditransitive dative variant*
[The waiter]$_{subject}$ [gave]$_{verb}$ [my cousin]$_{recipient}$ [some pizza]$_{theme}$

b. *The prepositional dative variant*
[The waiter]$_{subject}$ [gave]$_{verb}$ [some pizza]$_{theme}$ [to my cousin]$_{recipient}$

We know that the prepositional dative appeared alongside the ditransitive dative in late Old English (Brinton & Arnovick 2017:300), and we also understand the major probabilistic factors that regulate the choice between the two dative variants (Bresnan et al. 2007; Röthlisberger et al. 2017; Szmrecsanyi et al. 2017). With that being said, we note that the literature has given much consideration to formal constraints (i.e., structural complexity) and to information-status related factors. Semantic factors, by contrast, have been given rather short shrift.

In the present paper, we wish to turn our attention to the influence semantic factors have on the choice of dative variant. However, rather than restricting ourselves to coarse-grained descriptors such as constituent animacy (Ransom 1979), we will explore the effect of the fine-grained semantic properties of the lexical material in the slots – recipient and theme – of a dative construction. In this endeavour, we propose a completely bottom-up approach to semantic annotation, by employing automatically-generated, corpus-based semantic predictors obtained from vector space models (VSMs, Lenci 2018). We will refer to these predictors as VSM-semantic predictors.

After discussing the background to this study in Section 2, we describe the methodological steps of the analysis (Section 3). Section 4, then, details the technicalities behind VSMs. In Section 5, we analyse traditional and VSM-semantic predictors statistically, and discuss the results of the analysis in Section 6. The discussion in Section 7 wraps up this study.

2. Background to the study

2.1. A probabilistic approach to the dative alternation

A syntactic alternation, such as the dative alternation, consists of two or more constructions, called variants, with a highly similar meaning, and these variants must be semantically/functionally interchangeable (In Labov's 1972:188 words, "Alternative ways of saying 'the same' thing"). From a cognitive perspective, contexts in which variants occur represent choice points for the individual speakers (Pijpops 2020:286); in other words, speakers need to decide between two or more constructions that share a relationship of near-synonymity or near-equivalence. From this definition emerges a view of syntactic variation hardly compatible with a categorical model of grammar, where language structure is governed by categorical, deterministic constraints. Rather, alternation research tends to be grounded in a usage-based, probabilistic model of grammar (Bybee and Hopper 2001). Within this probabilistic approach, a pioneering variationist study of the dative alternation is Bresnan et al. (2007). Over the years, their investigation of the dative alternation has engendered a growing body of literature (see chapter 2 of Röthlisberger (2018) for an extended literature review), with follow-up studies looking into variation across varieties of English (Szmrecsanyi et al. 2017) and in historical time (Wolk et al. 2013).

2.2. Measuring lexical effects in the dative alternation

To explore the probabilistic factors that constrain language users' dative choices, the traditional variationist approach has focused on high-level, top-down predictors, that is, descriptive categories of analysis defined a priori by the researcher. This is a very efficient way to encode grammatical features at sentence-level (such as relative length of themes and recipients, pronominality, definiteness), but when it comes to the semantic characteristics of the constituent slots in the alternation, only few, top-down properties are considered in the literature. The motivation for this 'lean' semantic coding is practical: annotating for semantics is labor-intensive, time-consuming, and challenging to perform objectively. Besides, annotation for top-down semantic properties is limited in the extent to which it can represent lexical-semantic richness and variation (Glynn 2014). In fact, when it comes to semantics the bulk of the literature restricts attention to the predictor (or semantic class) 'constituent animacy'.

One way of coming to terms with this 'lean' semantic coding is proposed in work by Röthlisberger et al. (2017), who include lexical random effects in regression analysis. Although this technique works reasonably well to increase the goodness of fit of regression models, it does not explain much in terms of influence of *generalizable* lexical-semantic characteristics of the constituent slots.

3. Methodology: from lexical effects to vectors and predictors

3.1. Modelling semantic factors using Vector Space Models

To understand how lexical-semantic factors might contribute to predicting dative variant choice, we modelled the distributional semantic representation of the head nouns – called *target words* in our study – in dative construction recipient and theme slots using Vector Space Models.

Vector Space Models, also called distributional models, are part of a larger research framework called distributional semantics (DS; Lenci 2018), whose core idea is that "difference of meaning correlates with difference of distribution" (Distributional Hypothesis; Harris 1954:156). This usage-based approach to meaning comprises various distributional techniques for the semantic representation of lexical content, all sharing the assumption that items that occur in similar contexts in a given corpus will be semantically similar, while those that occur in different contexts will be semantically different.

For the present analysis, we chose to employ a count-based type-level vector model, built with the python package Nephosem (QLVL, version 0.1.0). This model allows us to build vectors as an aggregation over all the attestations of a given lemma in a given corpus, taking the form of an overall numerical profile of the lemma. In doing so, the type-level vector abstracts away from the individual occurrences that realize the (potentially) polysemous meanings of a lexeme, encoding instead *'the patterns within the type-level matrices that are indicative of different senses'* of the lemma (Heylen et al. 2015:157).

3.2. The methodological workflow

The present study involves a series of methodological steps: steps 1–5 concern methodological preliminaries, steps 6–12 generate VSM-semantic predictors, and steps 13–16 cover the statistical analysis (section 4).

- Step 1: Ready a dataset about the alternation under study, where observations are annotated for various contextual characteristics (traditional predictors, such as e.g., constituent length)
- Step 2: From this dataset, extract the theme and recipient constituent head lemmas (target words)
- Step 3: Choose a training corpus to train the distributional semantic models
- Step 4: Check if the target words also occur in the training corpus wordlist: if so, lemmatize and annotate them with the appropriate part-of-speech tags; otherwise, exclude these target words from the analysis
- Step 5: Run a quality control check on the final target words list
- Step 6: Define parameter settings for the distributional models
- Step 7: From the training corpus, extract and count the raw co-occurrences of target words and linguistic context words

- Step 8: Create co-occurrence matrices, in which rows correspond to the target items and the columns to the context words
- Step 9: Weight each raw co-occurrence matrix using association strength measures
- Step 10: Compute the similarity matrices to represent the distributional distance between the target words, based on the weighted co-occurrence matrices
- Step 11: On the basis of the similarity matrices, cluster the target words with separate clusters for theme and recipient
- Step 12: Evaluate the distributional models obtained and their clusters, and choose the best model according to customary criteria
- Step 13: Perform a Conditional Random Forest (CRF) analysis to determine the explanatory importance of traditional and VSM-semantic predictors, separately and in combination
- Step 14: Perform a Binomial Logistic regression analysis with only traditional predictors, for the sake of teasing apart the power of lexical effects from other effects
- Step 15: Perform a Binomial Logistic regression analysis with only VSM-semantic predictors, for the sake of teasing apart the power of lexical effects from other effects
- Step 16: Compare dative alternation regression models with traditional predictors to dative alternation regression models with VSM-semantic predictors.

Preliminary to building VSMs (and the semantic predictors obtained from them) are the following steps: N = 2,380 theme and recipient head nouns were extracted from the pre-existing, manually annotated spoken US English dative alternation dataset in Bresnan et al. (2007), which is another way of saying that the dataset confers N = 1,190 dative constructions (steps 1–2). To model the distributional lexical context of our target words, the spoken subset of the Corpus of Contemporary American English (COCA) (Davies 2008; ~127 million words) was selected as a training corpus (step 3). In step 4, all dative attestations containing recipient and theme lemmas which did not occur in the training corpus COCA were excluded, resulting in a total number of N = 1,164 relevant dative occurrences. In line with the COCA part-of-speech (PoS) tagging standard, the resulting N = 590 recipient and theme head lemmas (i.e., our target words) were tagged using the CLAWS PoS tagger for English (Garside & Smith 1997). Ultimately, in step 5 we run a last quality check on the wordlist, converting the characters to lower case and manually checking for typos.

Four observations from our dataset, which contain theme and recipient target words, are presented in (2):

(2) *a.* DAT-4100
[if I]$_{\text{subject}}$ [gave]$_{\text{verb}}$ [it]$_{\text{theme}}$ [to the government]$_{\text{recipient}}$ they would just waste it.

b. DAT-4067
[The judge]$_{\text{subject}}$ [will usually, uh, give]$_{\text{verb}}$ [custody]$_{\text{theme}}$ [to the mother]$_{\text{recipient}}$ ninety-seven percent of the time.

c. DAT-2644
And so, you know, [we]$_{\text{subject}}$ ['ll give]$_{\text{verb}}$ [him]$_{\text{recipient}}$ [fifteen]$_{\text{theme}}$ which will teach him a lesson but it's not just, you know, horrible

d. DAT-2772
But [they]$_{\text{subject}}$ [give]$_{\text{verb}}$ the [guy]$_{\text{recipient}}$ [a job]$_{\text{theme}}$ in prison and make him pay his damn debt.

The next set of steps (6–12) creates the VSM-semantic predictors. Firstly, we define the parameter settings used to implement the models;[1] these include window-size of linguistic context and PoS filters, measures of semantic association, and the expected number of clusters to be used as semantic predictors (step 6).

In the steps 7 and 8, different large co-occurrence matrices are built. Each row represents a target word from the theme or recipient slot of each dative occurrence (e.g., the recipients *government*, *mother*, *him,* and *guy* in (2)), while each column represents a context word from the COCA corpus. The aggregation of the co-occurrence frequencies between a single target word and all the context words from the COCA constitutes a *word-type vector*, which is a distributional representation of a lemma in the corpus space.

Importantly, target words will only co-occur with a small number of context words, and the vector obtained will be a sparse vector. Accordingly, most of the co-occurrences of target and context words will result in a negative value. A customary practice in DS, therefore, is to transform, or better, weight the raw co-occurrence values using association strength measures such that the informative semantic relationships between words are brought to light (step 9). Table 1 shows

[1] For the sake of brevity, all technicalities relating to the construction of vector space models, including the parameters used to implement the models and the reproducible codes, can be found in the OSF repository: https://osf.io/6cdtm/?view_only=845427193625414aa5f7bc8d34d4966e

the weighted co-occurrence values using the popular, reinforced version of Pointwise Mutual Information (PMI), called Positive PMI (PPMI): the negative values are replaced by zero to better manage any inaccuracies resulting from them (Jurafsky & Martin 2020:109).

Table 1: Small, constructed example of PPMI weighted type-level vectors of the three target lemmas in (2). The columns represent the PPMI association value between each target and context words.

Target	daughter	Europe	it	dad	troop
government	0	2.1	0.4	0	1.74
mother	3.5	0	0.9	3.2	0
guy	1.5	0	1.2	0.3	0
him	0.2	0	4.1	0.3	0

From each co-occurrence matrix, a similarity matrix is subsequently computed to represent the distributional similarity between each target word, using pairwise distances such as cosine distance (step 10). The more the two target words are similar, the more the value will be close to 1; the more the two target words are different, the more the cosine value will be close to zero. In Table 2, we can observe how *mother* and *him* are closer in the vector space because of some context words shared in the co-occurrence matrix, such as *dad* and *it*. We would expect, then, to find *government* in a different part of the vector space.

Table 2: Constructed example of cosine-based similarity matrix of the three target lemmas in (2).

Target	government	mother	guy	him
government	1.0	0.0	0	0.05
mother	0.0	1.0	0.3	0.8
guy	0	0.3	1.0	0.8
him	0.05	0.8	0.8	1.0

In step 11, based on the resulting similarity matrices, theme and recipient heads were clustered, yielding groupings of semantically-related types constituting our bottom-up semantic predictors (see Section 6). Using the Partition Around Medoids (PAM) clustering algorithm (Kaufman & Rousseeuw 1990), implemented in R with *cluster::pam()* (Maechler et al. 2022), the central members of the clusters (called medoids) are identified in the data, after which the type-word vectors are grouped around them based on a similarity-distance metric.

Finally, in step 12 the best distributional model is selected, i.e., the model that best clusters and separates the types of senses. Each model is evaluated by computing (i) the Concordance-C index (C-value) [2] of its Conditional Random Forest, and (ii) the negative silhouette percentage (Rousseeuw 1987) for the tokens, i.e., the percentage of tokens assigned to the wrong cluster. We considered only models with a negative silhouette of less than 25% and a C-value higher than 0.75. Before selecting the best model, we also checked clustering consistency in the fifteen best models so as to rule out possible biases.

4. Traditional and VSM-semantic predictors compared

After having explained distributional models and VSM-semantic predictors, we now turn back to the traditional predictors for benchmarking purposes. Following Szmrecsanyi et al. (2017) and Bresnan et al. (2007), the following traditional predictors (which are already annotated in the pre-existing dataset) will be included in the benchmarking analysis:

- *Recipient/Theme.type:* pronominal (personal pronouns, demonstrative pronouns, impersonal pronouns) versus non-pronominal (noun phrase)
- *Recipient/Theme.definiteness:* definite (definite, definite proper noun) versus indefinite
- *Recipient/Theme.animacy:* animate (human and animal) versus inanimate (collective, temporal, locative, inanimate)
- *Semantics of the dative verb:* (i) transfer; (ii) communication; (iii) abstract
- *Length difference:* log difference of the length of the recipient and theme phrases in orthographically transcribed words (Bresnan & Ford 2010): *log*(Recipient.length) – *log*(Theme.length)

With these traditional predictors under our belt, we can now address two questions: (i) What is the importance of traditional and VSM-semantic predictors for predicting dative variant choice? (ii) Do these VSM-semantic predictors add to the explanatory power of the traditional predictors? To that end, we employ two variationist statistical analyses:

- To address the first question, we employ Conditional Random Forest (CRF) analysis (Tagliamonte & Baayen 2012). This multivariate statistical method enables us to assess and rank the predictors according

[2] The C-value is non-parametric measure of how well a statistical model fits a set of observations, where C = 1 corresponds to optimal model prediction and C < = 0.5 represents inability to predict.

to their explanatory importance (step 13).[3] Three CRFs were fitted: (i) with traditional predictors only, (ii) with VSM-semantic predictors only, (iii) with both sets of predictors[4].

- The last three steps of the analysis will address the second research question via regression analysis, by involving both traditional and VSM-semantic predictors which, on the basis of the previous CRFs analysis, turn out to have the most explanatory power. These predictors are input to a binary logistic regression analysis with mixed effects (in the *lme4* R package implementation; Bates et al. 2015) to assess the role of VSM-semantic predictors in predicting the alternation. We first fitted the model only with traditional predictors (step 14) and VSM-semantic predictors (step 15) separately. Next, we combined the most significant predictors of both models to assess if lexical factors can (subtly) influence the prediction when other factors are in play (step 16). We fitted the models using a manual backward selection of the predictors. For each model, we first fitted a maximal model including all the predictors as fixed effects, and the external-language factor *speaker* and the lexical effects *Theme and Recipient heads* as random effects (Röthlisberger et al. 2017). After a round of model optimization for multicollinearity and convergence issues, we pruned the models: first, we pooled infrequent random effect categories into 'other' categories (thresholds: 5 for *speaker* and 2 for *Theme and Recipient heads*). Subsequently, those predictors that lacked explanatory power, namely the ones with the highest p-value, were removed one by one. The resulting minimal regression models will be discussed in Section 6.3.

5. Results

5.1. VSM-semantic predictors

The combination of the different parameters (see Section 4, step 6) yielded 432 distributional models. The best model (step 12) contains fifteen clusters for each modelled slot, thirty in all, and a negative token silhouette of 0.016% for recipients and 0.06% for themes, as well as with a *C*-value of 0.95 for a conditional random

[3] This technique has rapidly become mainstream in variationist research because of its robustness: the prediction is based on a substantial number of ensembles of conditional inference trees built on subsets of randomly sampled data (see Levshina 2021 for an in-depth explanation).

[4] The CRF fitted in this step do not serve to select the best model. In step 12, we only look at the C-value of CRFs with only VSM-semantic predictors; here, we are interested in the predictive power of the statistical tool.

forest model predicting dative variant choice. Most of the parameter settings of the best model are also shared by the first ten models, ensuring the robustness of the clustering: 4-size context window, no vocabulary filter, dimensionality of 5000, PPMI as weighting measure, dimensionality reduction with 10 perplexity, and 15 clustering partitions.[5]

We performed an initial qualitative investigation of the predictors, by looking at the semantic consistency of the groupings and their plots in two-dimensional space using the visualization technique t-SNE (Van der Maaten & Hinton 2008) in its R version (*Rtsne*; Krijthe 2015).

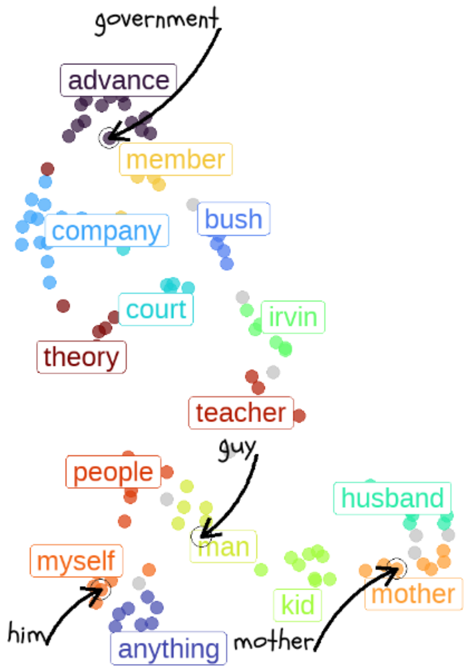

Figure 1: Scatterplot of recipient clusters. The clusters are labelled according to their central medoid.

[5] See the supplementary material in the OSF folder for more detailed information.

The predictors for the recipients share a noticeably clear pattern of consistent grouping (Figure 1): at the top-left of the scatterplot, we observe the type *government* in the cluster groupings related to economics, job titles, and law terminology. Below, we can observe a large cloud of clusters containing lemmas referring to humans and anaphoric pronouns such as *him* and *guy*. The small cloud of clusters at the bottom-right contains lemmas related to family roles, and the target-word *mother*.

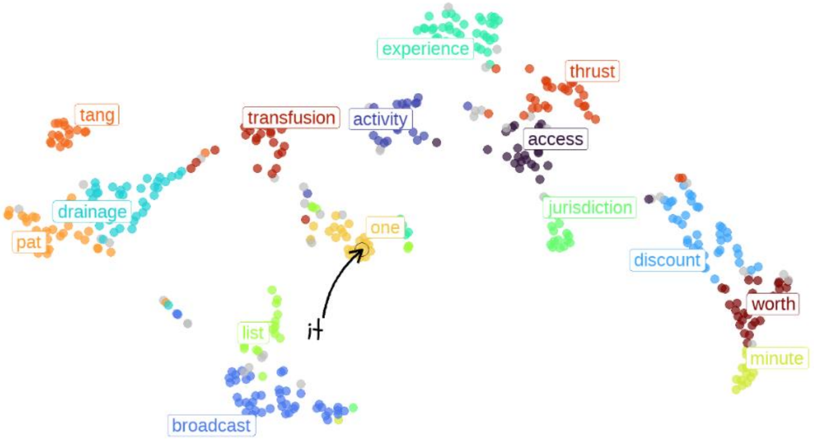

Figure 2: Scatterplot of theme clusters. The clusters are labelled according to their central medoid.

A similar pattern, but sparser, is presented by the theme clusters (Figure 2): household items are grouped together in the top-left corner, whereas a cloud of clusters mostly containing abstract nouns can be observed in the middle of the scatterplot. In the right-hand corner, clusters of lemmas related to the labor market are grouped together. The cluster *one* is mostly populated by grammatical markers, such as the extremely frequent pronoun *it.*

5.2. Most important predictors with CRF

The CRF plot in Figure 3 shows the importance ranking of the predictors, both VSM-semantic and traditional.

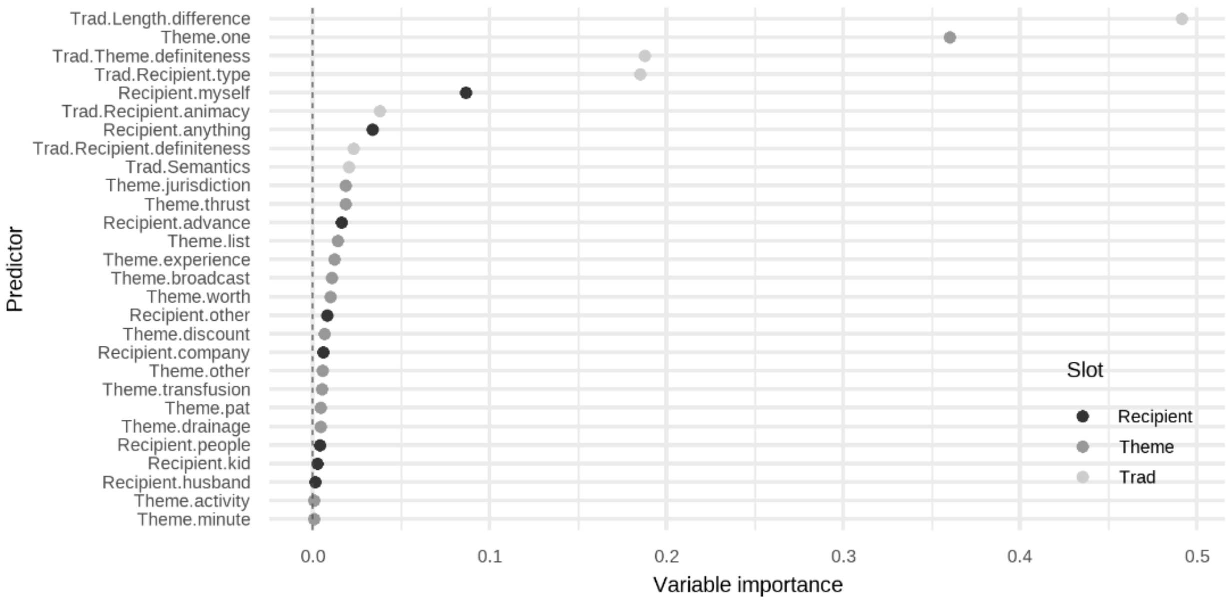

Figure 3: Conditional random forest.

As already reported in Szmrecsanyi et al. (2017), the traditional predictor 'length difference between recipient and theme' plays a key role in determining the choice between variants in US English, together with theme definiteness and recipient pronominality. That said, some VSM-semantic predictors rank highly as well: the cluster (and hence, VSM-semantic predictor) *Theme.one*, dominated by the pronoun *it*, follows the 'length difference' predictor in importance. *Recipient.myself*, a cluster mainly composed of pronouns and anaphorical markers, occupies the rank right below the traditional predictor 'recipient pronominality', followed distantly by *Recipient.anything,* a similarly composed grouping.

A large number of theme clusters also have a predictive effect, even if slight, as well as traditional predictors 'recipient animacy' and 'semantics of the verb' .

5.3. Regression modeling

As pointed out in Section 4, regression allows us to determine the importance of fine-grained semantic properties of the theme and recipient nouns for predicting variant choice.

Specifically, the C-values of the model with only traditional predictors and the one combining traditional and VSM-semantic predictors provide insight into how the classic traditional models would benefit from employing VSM-semantic predictors (Table 3). The results confirm that goodness of fit of models can be improved positively, although marginally, by lexical random effects, as already shown by Röthlisberger (2018). However, even though the model with the mixed predictors (i.e., traditional and VSM-semantic) yields the highest C-value with the random effects improvement (0.995), the high C-value of 0.96 of the traditional model without random effects suggests that traditional factors perform better alone than in combination with the VSM-semantic ones.

Table 3: C-values of logistic regression models with traditional only and mixed (traditional and VSM-semantic) predictors

	RM only traditional predictors	RM traditional and VSM-semantic predictors
C-value fixed effects	0.990	0.995
C-value fixed and random effects	0.957	0.937

6. Discussion

The present study has explored a new way of accounting for and measuring the effect of fine-grained lexical-semantic distinctions on the choice of the dative variant in English. We accomplished this task by relying on distributional semantics techniques, and restricted attention to the meaning of the materials in the constituent slots of the dative alternation.

We saw that some VSM-semantic clusters are highly predictive of the alternation, especially clusters containing anaphorical markers as pronouns. We also demonstrate how VSM-semantic and traditional predictors complement each other. At the same time, regression analyses indicated that traditional predictors outperform VSM-semantic predictors in terms of model performance, which is based on the C-value.

It should be kept in mind that this is a preliminary study, and we stress that the high frequency of pronouns and anaphorical markers in both the recipient and theme slots (as already noticed by Bresnan et al. 2007:89) needs further consideration. In follow-up research, we will therefore explore other alternations that might yield semantically more interesting results, such as the clausal complementation alternation in the history of English (3), and the progressive alternation in Italian (4):

(3) a. *The gerundial -ing clause variant*
[The boys]$_{\text{subject}}$ [remembered]$_{\text{matrix verb}}$ having [left]$_{\text{complement verb}}$ the party

b. *The that-clause variant*
[The boys]$_{\text{subject}}$ [remembered]$_{\text{matrix verb}}$ that they had [left]$_{\text{complement verb}}$ the party

(4) a. *The gerundial clause variant*
[L'insegnante]$_{\text{subject}}$ sta [correggendo]$_{\text{main verb}}$ i testi
The teacher stays correcting the texts
'the teacher is correcting the texts'

b. *The indicative mood clause variant*
[L'insegnante]$_{\text{subject}}$ [corregge]$_{\text{main verb}}$ i testi
The teacher corrects the texts
'the teacher corrects/is correcting the texts'

Additionally, we will experiment with token-level modelling of semantic space (see Montes 2021), which would enable us to conduct a more in-depth analysis of the correlation of individual variant instantiations with the context. This is extremely useful when lemmas are polysemous. Our ultimate goal is to assess the role of lexical-semantic properties of the embedding lexical context more fully for predicting variant choice, and to establish how this new knowledge can be combined with traditional models to advance theorizing about linguistic variation.

7. References

Aissen, J. (2003). 'Differential object marking: Iconicity vs. economy'. *Natural Language & Linguistic Theory*, 435-483.

Bates D, Mächler M, Bolker B, Walker S (2015). 'Fitting Linear Mixed-Effects Models Using lme4'. *Journal of Statistical Software, 67(1),* 1–48.

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). 'Predicting the dative alternation'. *Cognitive Foundations of Interpretation*, 69-94. KNAW.

Brinton, Laurel J. and Arnovick, Leslie K. (2017). *The English Language : A Linguistic History. Third ed.* Don Mills Ontario Canada: Oxford University Press.

Bybee, J., & Hopper, P. (2001). 'Introduction to frequency and the emergence of linguistic structure'. *Typological Studies in Language*, *45*, 1-26.

Collins, P. (1995). 'The indirect object construction in English: An informational approach'. *Linguistics 33*: 35-49.

Davies, Mark. (2008-) *The Corpus of Contemporary American English (COCA)*. Available online at https://www.english-corpora.org/coca/.

Garside, R., and Smith, N. (1997) 'A hybrid grammatical tagger: CLAWS4'. In Garside, R., Leech, G., and McEnery, A. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, London, pp. 102-121.

Glynn, D. (2014). 'The many uses of *run*: Corpus methods and Socio-Cognitive Semantics'. In D. Glynn and J. A. Robinson, eds., *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*, Amsterdam; Philadelphia: John Benjamins Publishing Company, 117–144.

Harris, Z. S. (1954). 'Distributional structure'. *Word, 10(2-3)*, 146-162.

Heylen, K., Wielfaert, T., Speelman, D., & Geeraerts, D. (2015). 'Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis'. *Lingua*, 157, 153-172.

Jones, M. N., Willits, J., & Dennis, S. (2015). 'Models of semantic memory'. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *The Oxford Handbook of Computational and Mathematical Psychology* (pp. 232–254). Oxford University Press.

Jurafsky, D., and Martin, J. H. (2020). *Speech and Language Processing* (3rd ed.).

Kaufman, L., & Rousseeuw, P. J. (1990). 'Partitioning around medoids (program pam)'. *Finding Groups in Data: an Introduction to Cluster Analysis*, 344, 68-125.

Krijthe, J. H. (2015). 'Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation'. R package version 0.16, https://github.com/jkrijthe/Rtsne.

Labov, W. (1972). *Sociolinguistic Patterns (No. 4).* University of Pennsylvania press.

Lenci, A. (2018). 'Distributional Models of Word Meaning'. *Annual Review of Linguistics*, 4(1), 151–171.

Levshina, N. (2021). 'Conditional inference trees and random forests'. In: *A Practical Handbook of Corpus Linguistics*, 611-643. Cham: Springer International Publishing.

Maechler, M., Rousseeuw, P., Struyf, A., Huber, M., Hornik, K. (2022). 'cluster: Cluster Analysis Basics and Extensions'. R package version 2.1.4

McInnes, L., Healy, J., & Melville, J. (2018). 'Umap: Uniform manifold approximation and projection for dimension reduction'. arXiv preprint arXiv:1802.03426.

McRae, K., & Jones, M. (2013). *14 Semantic Memory* (Vol. 206). Oxford: Oxford University Press.

Montes, M. (2021). *Cloudspotting: Visual Analytics for Distributional Semantics*. PhD Thesis.

Pijpops, D. (2020). 'What is an alternation? Six answers'. *Belgian Journal of Linguistics*, *34*(1), 283-294.

QLVL. (2021). *Nephosem (0.1.0)*. Zenodo. https://doi.org/10.5281/zenodo.5710426

Ransom, E.N. (1979). 'Definiteness and Animacy Constraints on Passive and Double-Object Constructions in English'. *Glossa, 13*, 215-240.

Röthlisberger, M. (2018). *Regional Variation in Probabilistic Grammars. A Multifactorial Study of the English Dative Alternation*. PhD Thesis.

Röthlisberger, M., Grafmiller, J., & Szmrecsanyi, B. (2017). 'Cognitive indigenization effects in the English dative alternation'. *Cognitive Linguistics*, *28*(4), 673-710.

Rousseeuw, P. J. (1987). 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis'. *Journal of computational and applied mathematics*, *20*, 53-65.

Szmrecsanyi, B., Grafmiller, J., Bresnan, J., Rosenbach, A., Tagliamonte, S., & Todd, S. (2017). 'Spoken syntax in a comparative perspective: The dative and genitive alternation in varieties of English'. *Glossa: a Journal of General Linguistics*, 2(1).

Tagliamonte, S. A., & Baayen, R. H. (2012). ‘Models, forests, and trees of York English: Was/were variation as a case study for statistical practice’. *Language Variation and Change*, *24*(2), 135-178.

Van der Maaten, L., & Hinton, G. (2008). ‘Visualizing data using t-SNE’. *Journal of Machine Learning Research*, *9*(11).

Wolk, C., Bresnan, J., Rosenbach, A., and Szmrecsanyi, B. (2013). ‘Dative and Genitive Variability in Late Modern English: Exploring Cross-Constructional Variation and Change’. *Diachronica, 30(3),* 382–419.