Benedikt Szmrecsanyi and Alexandra Engel

Register Variation in a Cognitive (Socio)linguistics Perspective

Abstract: Key questions in Cognitive (Socio)Linguistics include the following: “How do language users acquire lectal competence, how is it stored mentally, and how does it work in language production?” (Geeraerts, Kristiansen & Peirsman 2010: 10). We aim to shed more light on the storage and production component of this question. Specifically, we will explore the extent to which language users have different probabilistic grammars for different situational varieties of speech and writing (“registers”) – do our linguistic choice making processes differ depending on whether we engage in e.g., informal conversation or write blog entries? This issue is under-researched but loaded theoretically. Our case study is about the dative alternation in English (John gave the president a present versus John gave a present to the president). The methodology is usage-based and relies on both corpus evidence (i.e., observation) and a rating task experiment. We distinguish between four broad registers: informal spoken language, formal spoken language, informal written language, and formal written language. Analysis shows that different registers do indeed come with different probabilistic grammars, which indicates that lectal/register differences play an important role in cognitive categorization.

Keywords: register, probabilistic grammar, variation, dative alternation

Acknowledgement: This paper sketches a four-year project funded by the Research Foundation – Flanders (FWO) (grant # G0D4618N).

1 Introduction

This paper is about how the way people choose between “alternate ways of saying ‘the same’ thing” (Labov 1972: 188) depends on the situation of spoken and written language production, a.k.a. register. We stress that register variation as one important manifestation of lectal variation (including the distinction

Benedikt Szmrecsanyi & Alexandra Engel, KU Leuven, Department of Linguistics, Blijde-Inkomststraat 21, PO Box 03308, B-3000 Leuven, Belgium, e-mail: {benszm, alexandra.engel} @kuleuven.be

https://doi.org/10.1515/9783110733945-31

between formal and informal text types) has been a key focus in Leuven-school lectometry and cognitive sociolinguistics (see e.g., Speelman, Grondelaers and Geeraerts 2006). Our point of departure then is that register variation is rampant in human language (Ferguson 1983: 154), and that knowledge of how to use language in particular situations is a key ingredient of language users' lectal knowledge.

2 State-of-the Art

Previous research on register variation has primarily focused on the text frequencies of particular linguistic features in particular registers: how often or rarely do we find particular linguistic features, such as passive constructions, in particular registers? The flagship method in this line of research is the Multi-Dimensional (MD) approach developed by Douglas Biber (1988), which measures co-occurrence patterns of linguistic features. An alternative, variationist (in the spirit of, e.g., Labov 1972; Grondelaers and Speelman 2007) way of approaching register variation does not ask how frequently particular features are used in particular registers, but the following question instead: when speakers can choose between different ways of saying the same thing, what is the extent to which they draw on different choice-making processes in different registers? Such probabilistic register differences have received short shrift in the past. Variationist sociolinguists in the tradition of Labov's work would have the methodological toolkit to investigate these issues, but this community happens to be mostly interested in one particular register, vernacular speech as observable in sociolinguistic interviews (but see e.g., D'Arcy and Tagliamonte 2015). Probabilistic effects also take center stage in Probabilistic Grammar work à la Joan Bresnan and collaborators, but again most of the extant work in this tradition is concerned with spoken language (exceptions include e.g., Bresnan et al. 2007, Grafmiller 2014). In sum, the register-sensitivity of probabilistic choice-making should be of central theoretical importance to analysts working in experienced-based and usage-based paradigms, but so far this sensitivity is ill-understood and in want of empirical investigation.

3 Research Questions

This paper conducts a pilot study for the sake of determining the extent to which language users' probabilistic grammars may include knowledge about lectal register differences. The goal is to investigate the degree to which language users' choice-making processes are different as a function of register. We thus aim to assess – via corpus analysis and rating task experiments – how the effect size and direction of language-internal constraints on variation interacts with register as a language-external parameter. Two more specific research questions guide our analysis in the present paper:

1. With regard to register distinctions – what are the relevant register-related dimensions of variability: formal versus informal (formality), or written versus spoken (medium)?
2. With regard to probabilistic constraints – what are the probabilistic constraints that tend to have particularly variable probabilistic effects across registers?

4 Methodology

The alternation we analyze by way of a case study is the dative alternation after the verb *give* in English. To encode dative relations, speakers and writers of English may use two semantically roughly equivalent structural patterns: the ditransitive dative variant, as in (1a), and the prepositional dative variant, as in (1b):

(1) a. *Several charities have different stances on whether or not you should give* [*homeless people*]$_{\text{recipient}}$ [*money*]$_{\text{theme}}$ *directly*
(The Independent, 10/01/2018)
b. *Mm and I used to give* [*a lot of money*]$_{\text{theme}}$ [*to homeless groups*]$_{\text{recipient}}$
(BNC2014, SPHJ)

The dative alternation is one of the best-understood alternations in the grammar of English. One seminal study on the dative alternation in English is Bresnan et al. (2007), which explores factors that constrain language users' dative choices in American English (telephone) conversations. Bresnan et al. find that variation between the two dative options is constrained by about 10 predictors, including e.g., pronominality of the recipient/theme, discourse accessibility (pragmatics), and animacy of the recipient. If, for example, unlike in (1) the recipient is inanimate, Bresnan et al. 's regression model predicts that the odds for the prepositional dative increase by a factor of about 4. This is the probabilistic effect that inanimate recipients have on dative choice in telephone conversations. But do inanimate recipients have the same effect in, say, formal speeches? What about the other predictors? What is the extent to which language users have to adjust probabilistic decision-making when they switch from telephone conversations to other registers? These are the kinds of questions that we are interested in.

4.1 Corpus-based track

A corpus-based variationist analysis applying logistic regression to a richly annotated dataset was carried out (see Szmrecsanyi 2019 for discussion). We chose four registers at the intersection between formality and mode. Focusing on British English, we selected the Spoken BNC2014 (~11.4 million words) for informal conversations between friends and family members (Love et al. 2017); a corpus of Hansard transcriptions from House of Commons debates for formal spoken language (~59.4 million words) (Marx and Schuth 2010); the British English blogs part of the GloWbE corpus for informal written language (~148 million words) (Davies 2013); and a corpus of newspaper articles from *The Independent* (~113.5 million words) representing formal written language (JSI Newsfeed corpus, Bušta et al.).

From these corpora, we automatically extracted tokens of the verb *give*, which were then manually checked for criteria of the variable context. Accordingly, tokens were filtered out which included only one constituent, mistaggings (e.g., *given* as preposition or adjective), non-canonical word order, clausal constituents, *give* as particle verb, fixed expressions, passivized or relativized constructions, and constructions in which the *to*-phrase depended on the theme (as in *give the answer to that question*). For reasons of speaker/author contribution, direct quotes were also filtered out. From the remaining tokens, a balanced, random sample of 2,600 observations was created (i.e., 650 tokens per corpus, half of which were ditransitive dative constructions and the other half were prepositional dative constructions). The dataset was annotated for the following predictors: pronominality (pronominal vs. non-pronominal), animacy (animate vs. inanimate), definiteness (definite vs. indefinite), constituent length (in number of characters), complexity (simple vs. complex), and (head) frequency of both constituents as well as verb sense (transfer of concrete object, transfer of abstract object, communication sense).[1] Constituent length measures were combined into a single predictor, Weight Ratio, by dividing recipient length by theme length.

[1] Collective nouns were annotated as ‘inanimate’. Complex constituents included restrictive postmodifications to the constituent’s head.

Numerical predictors were log-transformed and standardized to reduce multicollinearity.

A logistic mixed effects regression model was then fitted in *R* using the *lme4* package (Bates et al. 2015). To test for the effect of register on the internal constraints, three interactions between Register and Weight Ratio, Recipient Definiteness, and Theme Definiteness were included in the model in addition to the main effects for all of the above predictors. All levels were set to the default levels of the ditransitive dative. Random effects for recipient and theme head lemma as well as for speaker identity account for idiosyncrasies. Model selection followed a backward elimination process. The resulting model has a high *C* index of 0.97, confirming outstanding model performance, and an accuracy of 91.4% (baseline 50%).

4.2 Experimental track

This corpus model was then tested against human rating performance. Are language users sensitive to probabilistic patterns in the choice of dative variants? More specifically, do we find similar patterns in a comparison between corpus-based predictions and language users' intuitions about the probability of variants? To this end, we set up a rating task experiment in which we presented participants with both variants in authentic corpus examples. Participants were asked to give gradient ratings as to which variant they judge more likely in the context. Previous research in this vein has found converging evidence between corpus results and ratings (Bresnan and Ford 2010; see Klavan and Divjak 2016 for a review). In a seminal study on the dative alternation, Bresnan and Ford (2010) asked American English and Australian English speakers to distribute 100 points across both variants and found variety-specific probabilistic effects corresponding to the patterns found in the corpus. Do we find similar patterns when we examine register-specific knowledge?

Based on the corpus results, spoken informal and spoken formal registers were chosen to provide the items for the present experiment (see Section 5.1). In total, 32 corpus excerpts were selected, 16 for each register. Per register, six items contained dative constructions and ten filler items contained either lexical, register-specific choices (four items) or the choice between the relativizers *which* and *that* (six items). These fillers were included to distract from the target constructions. Target items came from six probability bins across the whole probability range (probability of 0.06-0.99 for the prepositional dative). In order to control for possible confounding variables, the target items were restricted to those items that included simple, non-pronominal constituents and a definite recipient. As a result, three items per register were not included in the dataset and were thus unseen by the corpus model. Per register, target items were counterbalanced for dative variant, theme definiteness and whether they were seen or unseen by the corpus model. The full item set will be published as part of Engel et al. (in preparation).

Two lists were created by varying the presentation side of the original variant. All items of one register were presented after one another followed by all items of the other register. Per list, there were two versions to account for possible order effects: version A began with the formal spoken items, version B with the informal spoken items. Eight simple comprehension questions were included to ensure that participants read the excerpts carefully.

The rating task was implemented in an online survey using Qualtrics. Participants were sampled via Qualtrics Research Services. One hundred British English monolingual native speakers (50 female, 50 male; mean age: 55 years old; age range: 19-78) took part in the study and gave informed written consent before completing the survey. Participants were asked to indicate their ratings by means of a slider bar. Mean duration was 26 minutes.

For the analysis, eleven participants were excluded either due to low accuracy (< 75% correct answers) in response to the comprehension questions or due to excessive time spent on the survey (> 40 minutes; based on interquartile range). Ratings were standardized and entered as dependent variable in a linear mixed effects regression model with Predicted Corpus Probability and Weight Ratio, and an interaction between Register and Theme Definiteness as explanatory variables. In addition, a random effect of participant with Predicted Corpus Probability in the slope was added to account for participant-specific variability.

5 Results

5.1 Corpus-based track

The corpus model indicates that the prepositional dative becomes more likely when the recipient is inanimate ($\beta = 0.95$, $p <.001$), indefinite ($\beta = 2.28$, $p <.001$), non-pronominal ($\beta = 2.18$, $p <.001$), complex ($\beta = 0.68$, $p = .02$), longer than the theme ($\beta = 1.84$, $p <.001$) and when the theme is simple ($\beta = 2.19$, $p <.001$), definite ($\beta = 1.06$, $p = .005$), and pronominal ($\beta = 2.18$, $p = .005$). Main effects for Verb Sense and Register indicate that the prepositional dative becomes more likely when *give* has a communication sense ($\beta = 1.37$, $p < .001$) or a transfer sense ($\beta = 0.73$, $p = .011$), and that the probability of a prepositional dative with *give* is higher in spoken formal ($\beta = 1.06$, $p = .005$) and written informal ($\beta = 0.96$, $p = .008$) registers compared to the spoken informal register. Moreover, there are interactions between Register and Recipient Definiteness (Figure 1) and between Register and Theme Definiteness (Figure 2). These interactions show that the effect size of Recipient Definiteness is modulated in the spoken formal register compared to the spoken informal register ($\beta = -2.02$, $p = .003$) and that the direction of the effect of Theme Definiteness is reversed in the spoken formal register ($\beta = -1.7$, $p = .001$). The random effect for theme head lemma significantly contributes to explaining the variation ($\sigma^2 = 3.27$).

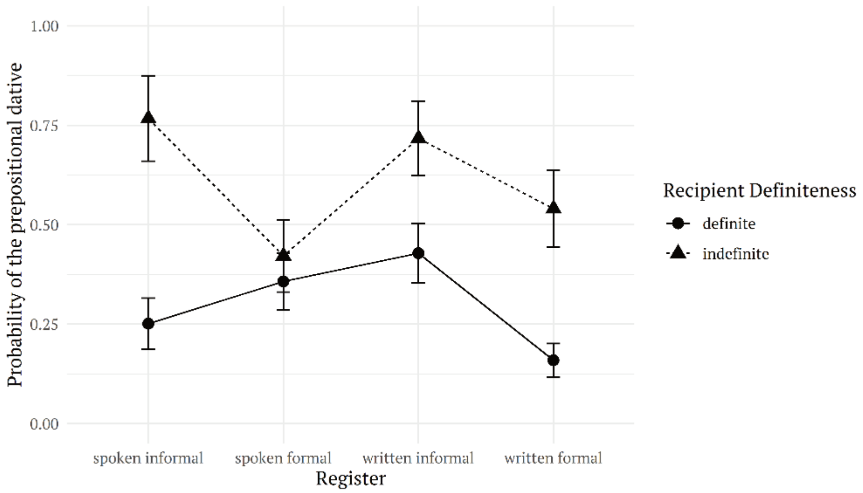

Fig. 1: Interaction effect between register and recipient definiteness in corpus model. The probability of the prepositional dative (y-axis) increases when the recipient is indefinite across all registers (x-axis), but the magnitude of the effect is modulated in the spoken formal register. Error bars represent standard errors.

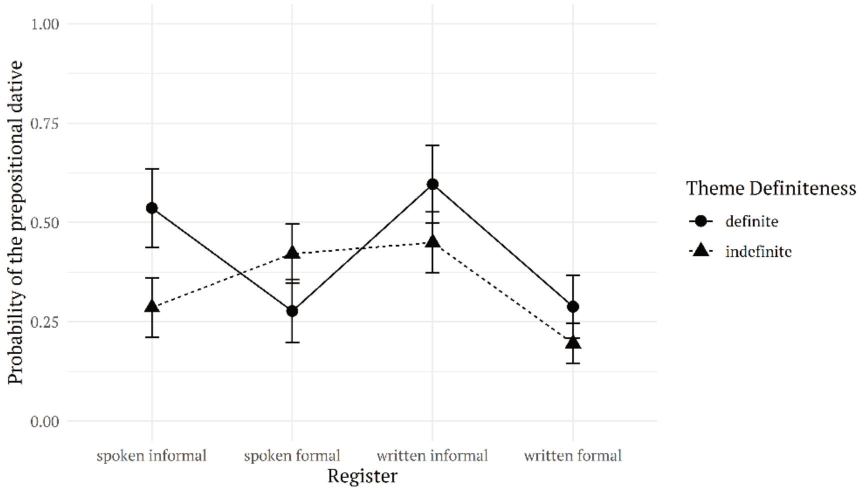

Fig. 2: Interaction effect between register and theme definiteness in corpus model. The probability of the prepositional dative (y-axis) increases when the theme is definite except in the spoken formal register where the direction of the effect is reversed. Error bars represent standard errors.

5.2 Experimental track

There is a main effect for Predicted Corpus Probability ($\beta = 0.3$, $p < .001$), indicating that participants gave higher ratings for the prepositional dative as the predicted probability for the prepositional dative in the corpus model increases. In addition, there is an interaction between Register and Theme Definiteness ($\beta = 0.39$, $p = .001$; see Figure 3). Participants gave higher ratings for the prepositional dative in spoken formal items with indefinite themes. These results show that the corpus model and participants' ratings pattern together. There were also main effects for Weight Ratio ($\beta = -0.19$, $p <.001$), and Theme Definiteness ($\beta = -0.33$, $p < .001$). Note that with a conditional R^2 of 0.11 and a marginal R^2 of 0.09, the model leaves a large part of the variance unexplained, which might be due to individual variation in the extent to which participants made use of the rating scale.

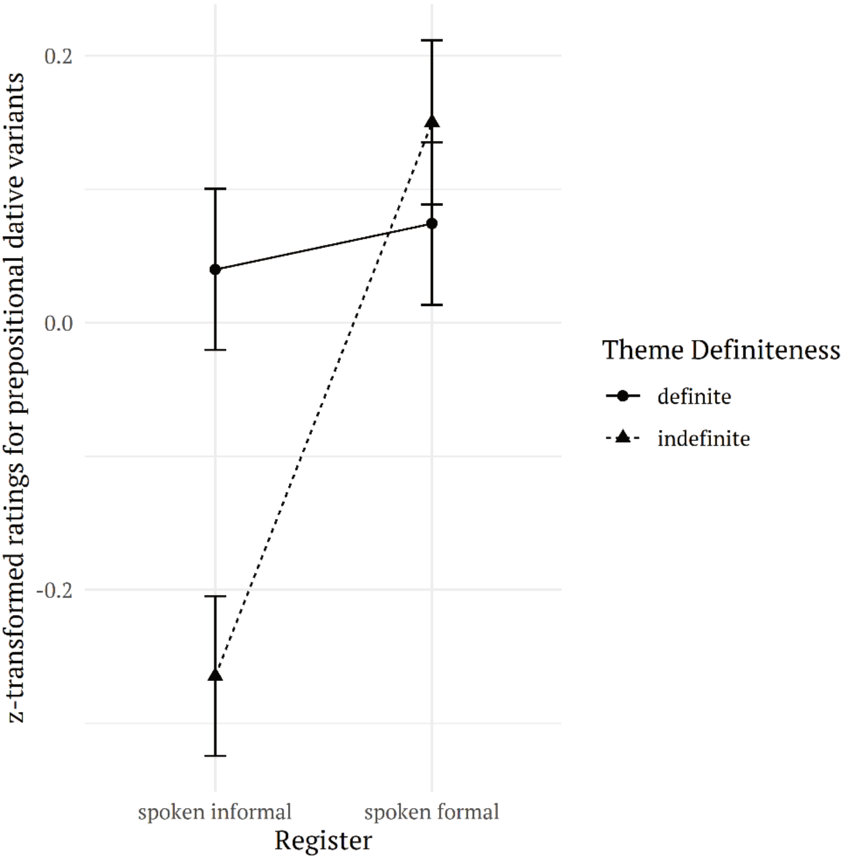

Fig. 3: Interaction between register and theme definiteness in participants' responses. Participants gave higher ratings for the prepositional dative in spoken informal items when the theme was definite in contrast to higher ratings for the prepositional dative in spoken formal items with indefinite themes. Ratings are expressed as z-scores.

Results for the filler items show an interaction between Register and Filler Type ($\beta = -0.39$; $p = 0.002$). In the spoken formal register, participants had stronger preferences for the formal variant in lexical items compared to relativizer items. Additionally, a main effect for Register ($\beta = 0.9$; $p < .001$) indicates that formal variants overall received higher ratings in the formal register. These results confirm that participants made register-specific judgments.

6 Discussion

Analysis shows that the main effects are in line with “harmonic alignment” (Bresnan et al. 2007; Bresnan and Hay 2008; Theijssen et al. 2013; Röthlisberger, Grafmiller and Szmrecsanyi 2017 and/or “Easy First” effects (MacDonald 2013): early constituents tend to be simple, short, animate, and definite. But what about interactions with register? According to corpus data, register interacts with the effect of definiteness:

- In all registers, the prepositional dative is more likely when the recipient is indefinite, but the largest definiteness effect can be observed in the spoken informal register, while we find the smallest effect in the spoken formal register.
- As to the theme, the prepositional dative is overall more likely when the theme is definite, but the direction of the effect is reversed in the spoken formal register. As to effect sizes, we observe the largest effect in the spoken informal register, and the smallest effects in both formal registers.

The experimental analysis partially confirms the existence of these differences: in the rating data as well, we see an interaction between register and theme definiteness. Register-specific effects are subtle, but subjects still seem to be sensitive to such subtle effects. That said, there is a great deal of individual variation.

Why does definiteness interact with register? Supplementary analysis demonstrates that indefinite recipients are particularly frequent in the ditransitive dative in the spoken formal register. Assuming that definite referents are more accessible than indefinite ones (Gundel, Hedberg and Zacharski 2001), we may argue that in spontaneous conversation, definite referents are placed first because they are easier to access and to process. We also find more indefinite themes in the spoken formal register overall, and with the prepositional dative in particular; it thus seems that in general, more indefinite referents are used in parliamentary debates compared to informal conversations. This might be explained by the high frequency of definite pronouns in informal conversations, as opposed to the higher frequency of nouns in more informational registers (Biber et al. 1999: 235).

We now move on to a discussion of the wider significance of these results. Our findings have implications and relevance for theory formation. Our research is ultimately concerned with the nature and scope of linguistic knowledge, and with the interaction of this knowledge with socioculture (for register conventions are social in nature). Generally speaking, variationist sociolinguists believe that "internal constraints [...] are normally independent of social and stylistic factors" (Labov 2010: 265), and it is of course this independence that our findings call into question. Given that definiteness as a probabilistic constraint has demonstrably different effect sizes (and sometimes even effect directions) across registers, Guy's Grammatical Difference Hypothesis (Guy 2015), according to which having different constraints means having different grammars, would arguably warrant us to conclude that language users have a number of different register-specific grammars, akin to situations of diglossia or bilingualism. So, coming back to the cognitive sociolinguistics research question spelled out in the abstract – How is lectal competence stored mentally, and how does it work in language production? (Geeraerts, Kristiansen and Peirsman 2010: 10) – our analysis would seem to suggest that competence about register differences is maybe more crucial and, in fact, richer than previously assumed by many: if different register come with different (probabilistic) grammars as we have shown, then register competence is no different from multilingual or multidialectal competence.

References

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1).

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Pearson Education Limited.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina & Harald Baayen. 2007. Predicting the dative alternation. In G. Bouma, I. Kraemer & J. Zwarts (eds.), *Cognitive foundations of interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bresnan, Joan & Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86(1). 168–213.

Bresnan, Joan & Jennifer Hay. 2008. Gradient grammar: An effect of animacy on the syntax of give in New Zealand and American English. *Lingua* 118(2). 245–259.

Bušta, Jan, Ondřej Herman, Miloš Jakubíček, Simon Krek & Blaž Novak. *JSI Newsfeed Corpus* (9th International Corpus Linguistics Conference). Birmingham.

D'Arcy, Alexandra & Sali A. Tagliamonte. 2015. Not always variable: Probing the vernacular grammar. *Language Variation and Change* 27(03). 255–285.

Davies, Mark. 2013. *Corpus of Global Web-based English: 1.9 billion words from speakers in 20 countries (GloWbE)*. https://www.english-corpora.org/glowbe/.

Engel, Alexandra, Jason Grafmiller, Laura Rosseel & Benedikt Szmrecsanyi. In preparation. Different registers, different grammars? A cognitive sociolinguistic study into register-specific effects in the English dative alternation.

Ferguson, Charles A. 1983. Sports announcer talk: Syntactic aspects of register variation. *Language in society* 12(02). 153–172.

Geeraerts, Dirk, Gitte Kristiansen & Yves Peirsman. 2010. Introduction. Advances in Cognitive Sociolinguistics. In Dirk Geeraerts, Gitte Kristiansen & Yves Peirsman (eds.), *Advances in Cognitive Sociolinguistics*, 1–20. Berlin, New York: De Gruyter Mouton. https://www.degruter.com/view/books/9783110226461/9783110226461.1/9783110226461.1.xml (accessed 7 December, 2020).

Grafmiller, Jason. 2014. Variation in English genitives across modality and genres. *English Language and Linguistics* 18(03). 471–496.

Grondelaers, Stefan & Dirk Speelman. 2007. A variationist account of constituent ordering in presentative sentences in Belgian Dutch. *Corpus Linguistics and Linguistic Theory* 3(2). 161-193.

Gundel, Jeanette K., Nancy Hedberg & Ron Zacharski. 2001. Definite descriptions and cognitive status in English: why accommodation is unnecessary. *English Language and Linguistics* 5(2). 273–295.

Guy, Gregory R. 2015. Coherence, constraints and quantities. In. Talk given at NWAV44, Toronto, date.

Klavan, Jane & Dagmar Divjak. 2016. The cognitive plausibility of statistical classification models: Comparing textual and behavioral evidence. *Folia Linguistica* 50(2). 355–384.

Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Philadelphia Press.

Labov, William. 2010. *Principles of linguistic change. Vol. 3: Cognitive and cultural factors* (Language in Society 39). Malden, Mass.: Wiley-Blackwell.

Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina & Tony McEnery. 2017. Compiling and analyzing the Spoken British National Corpus 2014. *International Journal of Corpus Linguistics* 22(3). 319–344.

MacDonald, Maryellen C. 2013. How language production shapes language form and comprehension. *Frontiers in psychology* 4. 226.

Marx, Maarten & Anne Schuth. 2010. DutchParl: The Parliamentary Documents in Dutch. In *Proceedings of the Seventh International Conference on Linguistic Resources (LREC-2010)*, 19–21. European Language Resources Association.

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

Speelman, Dirk, Stefan Grondelaers & Dirk Geeraerts. 2006. A profile-based calculation of region and register variation: the synchronic and diachronic status of the two main national varieties of Dutch. In Andrew Wilson, Dawn Archer & Paul Rayson (eds.), *Corpus linguistics around the world*, 181–194. Amsterdam-New York: Editions Rodopi B.V.

Szmrecsanyi, Benedikt. 2019. Register in variationist linguistics. *Register Studies* 1(1). 76–99.

Theijssen, Daphne, Louis ten Bosch, Lou Boves, Bert Cranen & Hans van Halteren. 2013.Choosing alternatives: Using Bayesian networks and memory-based learning to study the dative alternation. *Corpus Linguistics and Linguistic Theory* 9(2). 227–262.