How register-specific is probabilistic grammatical knowledge? A programmatic sketch and a case study on the dative alternation with *give*

Alexandra Engel[a], Jason Grafmiller[b], Laura Rosseel[c,a], Benedikt Szmrecsanyi[a], Freek Van de Velde[a]

[a] *KU Leuven, Leuven, Belgium*

[b] *University of Birmingham, Birmingham, United Kingdom*

[c] *Vrije Universiteit Brussel, Brussels, Belgium*

Running head: Register-specificity of probabilistic grammatical knowledge

Word count (incl. references, tables and figure captions): 10830

Abstract

While there is preliminary evidence about the importance of register in linguistic choice-making processes, systematic studies focusing on the interaction between register and language-internal constraints are lacking in variationist linguistics. This contribution sketches an ongoing project in which two well-understood grammatical alternations (dative alternation and future marker alternation) are analyzed with variationist methods, focusing on the role of register defined at the intersection of mode (spoken vs written) and formality (formal vs informal). Probabilistic corpus models will be complemented with rating experiments to investigate to what extent they correlate with participants' ratings, and to illustrate the importance of methodological diversity in investigating usage-based theories of grammar. We present corpus results of a case study on the dative alternation with *give*.

1. Introduction

Probabilistic grammars describe usage patterns of syntactic alternations as a function of quantifiable probabilistic constraints (see also Section 2). Register as predictor modulating language-internal constraints has been largely absent from studies into probabilistic effects in syntactic alternation studies. This contribution presents a programmatic sketch of an ongoing project at KU Leuven entitled *The register-specificity of probabilistic grammatical knowledge in English and Dutch* and discusses corpus results from a case study on the English dative alternation with *give* (see Section 4). In this project, we make use of variationist methods to elucidate how register as a language-external constraint influences the effect of language-internal constraints on the choice of a grammatical variant. To this end, we rely on customary text categories as registers, which we – in accordance with Biber & Conrad (2019: 6) – view as variation patterns associated with characteristics of the situational context of production in both speech and writing. Note, however, that our research design differs from text-linguistic designs that investigate the functional relationship between variation patterns and situational context (cf. infra).

A commonly applied text-linguistic method to study register variation is Multidimensional Analysis (MDA, Biber 1988), in which the goal is to describe register distinctions in terms of the functional relationship between linguistic variation (operationalized by rates of (co-)occurrence of linguistic features in

each text of a corpus) and characteristics of the situational context of the production circumstances (for a full description of the MDA approach, see Biber 1988; 2012; 2019; Biber & Conrad 2019). By contrast to text-linguistic approaches to register variation focusing on *frequencies* of occurrence of a set of linguistic features in a text (or sub-corpus), variationist methods are concerned with the *proportional* preferences for one variant over another (Biber 2012: Section 2; Biber et al. 2016: 357; see also Szmrecsanyi 2019). The unit of analysis are thus individual observations ('variants') of one single feature (i.e., 'variable') at a time (Szmrecsanyi 2019: 77) instead of texts. Since the object of study in variationist linguistics are functionally equivalent variants of a variable, the goal is to uncover probabilistic effects in choice-making processes between those variants.

The remainder of this chapter is structured as follows. Section 2 outlines variationist perspectives on register. Section 3 presents the methodology of the project. While we report only on corpus results here, the larger project investigates probabilistic grammatical knowledge by comparing corpus-based predictions to the performance of participants in rating task experiments (cf. Section 3.2). Section 4 presents results from a case study on the dative alternation with *give* in English, and Section 5 offers a discussion of these results and some concluding remarks.

2. Register in variationist linguistics

Variationist linguistics is concerned with linguistic 'variables' or "alternate ways of saying 'the same' thing" (Labov 1972: 188). Of particular interest is the probability of choosing one or the other variant choice given language-internal and language-external constraints (Tagliamonte 2013; Szmrecsanyi 2017). Language-internal constraints include, for instance, grammatical characteristics like animacy, pronominality, and definiteness. Language-external constraints include variety, a speaker's age or gender, and register. Comparative studies involving such language-external constraints may examine how language-internal constraints vary across different external sources of variability in three lines of evidence: (1) Which predictors are statistically significant? (2) What is the magnitude of the effect? (3) What is the order of levels, or constraint hierarchy, within a predictor? (Tagliamonte, 2013: 131). To answer these questions, a common approach within variationist linguistics is to fit logistic regression models on richly annotated datasets and explore interactions between language-external and language-internal constraints.

The notion of 'register' is "typically conceptualized as stylistic variation in aesthetic preferences" (Szmrecsanyi 2019: 78) pertaining to the situation of language production. The questions that variationist linguists pose are the

following: 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? And to what extent are probabilistic effects similar or different across registers? Hence, the main focus in variationist linguistics is on the probability of variant choice as opposed to text frequencies.

Put simply, we can distinguish two major strains of research that can be subsumed under the cover term 'variationist linguistics': variationist sociolinguistics and corpus-based variationist linguistics (cf. Szmrecsanyi 2017). Within variationist sociolinguistics in the tradition of William Labov, register has often been neglected as a language-external constraint in variationist linguistics, due to an assumption that "internal constraints are normally independent of social and stylistic factors" (Labov 2010: 265). In sociolinguistic theory, it has traditionally been assumed that there is a uniform and stable core grammar for each variety of language (Guy 2005: 562; but see D'Arcy & Tagliamonte 2015, for critical discussion). Traditional Labovian sociolinguistics has mainly focused on the "vernacular", an informal speech variety of everyday life, which is seen as the 'baseline' form of language since it is acquired first in life (Labov 1984: 27) and speakers pay "minimum attention" to it (Labov 1972: 208). Sociolinguistic interview corpora usually comprise only one register since their data is gathered in an attempt to capture the vernacular (Rickford 2014: 590). Therefore, variationist research has often

neglected situational and stylistic specificities of the variable grammar (Rickford 2014: 596). In corpus-based variationist linguistics, on the other hand, researchers usually rely on existing text categories as defined by corpus compilers, when register is included to study language variation. However, register is sometimes treated as "a nuisance factor" (Szmrecsanyi 2019: 77) as opposed to intrinsically interesting, which is why it should be controlled for in the analysis. For example, Gries (2015) recommends accounting for register-related idiosyncrasies by including register in a (nested) random effect structure in logistic mixed-effects regression models. Apart from studies including register as random effect (e.g., Heller et al. 2017; Ehmer & Rosemeyer 2018), other studies include register as main effect but not in interaction with other grammatical factors (e.g., Grondelaers et al. 2008; Geleyn 2017; Pijpops & Van de Velde 2018; Grafmiller & Szmrecsanyi 2018).

Despite this general neglect of register in variationist linguistics, previous research (in corpus-based variationist linguistics) indicates that register does play a role in the probabilistic effects in language when interaction effects are considered. Röthlisberger et al. (2017) investigated the dative alternation (as in *Tom gives Mary the book* vs *Tom gives the book to Mary*) in nine varieties of World Englishes and found that register interacts with variety, suggesting stylistic differences across varieties (see also Grafmiller & Szmrecsanyi 2018, on the particle placement alternation, e.g. *Sue picked up the*

book vs. *Sue picked the book up*). Previous studies on grammatical alternations in single varieties also found differences in constraints and the size of the effect of those constraints across different registers. For example, Theijssen et al. (2013) found that the effect of theme definiteness differs between spoken and written language in the British English dative alternation. Similarly, Grafmiller (2014) found substantial variation across six spoken and written genres for the effect sizes of language-internal factors governing the choice between the *s*-genitive and the *of*-genitive in American English. However, not all studies have found genre- or register-related differences in probabilistic choice-making (e.g., Tagliamonte 2016).[1]

With our project being situated in the corpus-based variationist linguistics paradigm, we aim to contribute to a better understanding of probabilistic register differences by focusing on the interaction between register and language-internal constraints. To this end, we draw inspiration from the Probabilistic Grammar framework developed by Joan Bresnan and colleagues (Bresnan et al. 2007; Bresnan & Hay 2008; Bresnan & Ford 2010), which makes three basic assumptions: First, grammatical variation is gradient and probabilistic in nature. Second, linguistic choice-making is conditioned by a

[1] Tagliamonte (2016) examined future time markers in three genres of written, computer-mediated youth language in Canadian English (e-mail, instant messaging, and SMS) and found that the constraints are stable across these genres, although the relative frequency of the future markers differs considerably from the ones found in the vernacular (cf. Tagliamonte & D'Arcy 2009).

multitude of probabilistically varying constraints. Third, language users have an internalized grammar that is sensitive to these probability patterns. Thus, this framework essentially adopts a usage-based approach in which linguistic knowledge is emergent from language use and linguistic experience (e.g., Bybee 2006).

In sum, we lack systematic investigations into the register-sensitivity of probabilistic choice-making which should be of central theoretical importance to analysts working in experienced-based and usage-based paradigms. Our project seeks to contribute to a better understanding of register from a corpus-based variationist perspective by specifically focusing on the interactions between language-internal constraints and register. With this research design we can shed more light on register variation from a different angle, complementary to text-linguistic approaches (Biber 2012; Biber et al. 2016).

3. A programmatic sketch

3.1 Research questions

We investigate the degree to which language users' choice-making processes are influenced by the stylistic demands of different registers. Thus, we do not

aim to identify and describe the features that distinguish registers from one another. Specifically, this project is designed to address four research questions:

RQ1: Where do we find most register-related variability with regard to probabilistic grammar – along the continuum of formality (formal vs. informal) or between modes (written vs. spoken)?

RQ2: What probabilistic constraints are particularly variable across registers?

RQ3: Are language users sensitive to register-specific probabilistic effects?

RQ4: Do closely related languages such as English and Dutch differ in terms of the importance of probabilistic register differences?

The first two questions are addressed in a corpus study in which we fit logistic regression models on richly annotated datasets in order to focus on the interactions between register and language-internal constraints (see Section 4.2). RQ1 and RQ2 will be addressed in the present paper by means of a case study on the dative alternation with *give*. RQ3 and RQ4 pertain to prospective research endeavors within the project. RQ3 will be addressed by conducting rating task experiments and ultimately correlating the results of those experiments with the corpus analyses. Additionally, we investigate alternations in two languages to address RQ4, namely, whether register differences can be found across languages. By doing so, we will be able to draw a more substantiated conclusion about the extent to which grammar varies

probabilistically across registers.

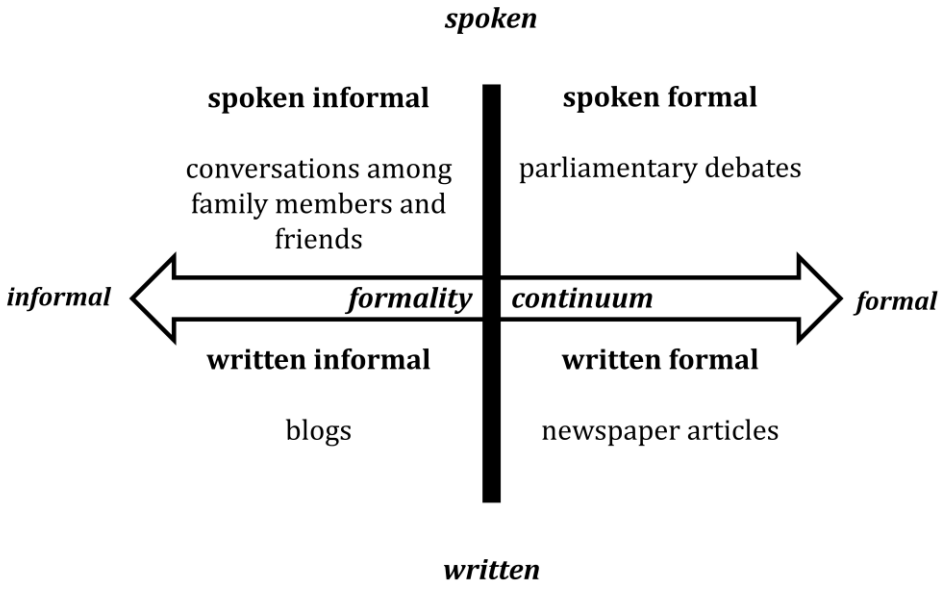
3.2 Methodology

Apart from tackling the need for variationist research involving register, an innovative aspect of this project is the combination of observations from language use and metalinguistic judgments of speakers and their predictive capacities. We use logistic regression to model the probability of a binary outcome (the choice of a variant) based on naturalistic language usage in corpus data. Predictions from our corpus models will then be tested against choice predictions made by speakers of the same variety in rating task experiments. This methodological diversity (as advocated by, e.g., Arppe et al. 2010; Schönefeld 2011) will allow us not only to tap into naturalistic production, but also into language processing and metalinguistic knowledge in order to be able to evaluate the cognitive reality of multivariate corpus models (Klavan & Divjak 2016). Converging evidence could possibly substantiate our conclusions drawn from these models. Even if the rating data is found to diverge from the corpus results, this could help us to shed light on differences in processes and factors involved in both language production (i.e., corpus data) and experimental data and advance our methodologies to study language use (Schönefeld 2011: 3f.; Arppe et al. 2010: 5).

3.2.1 Corpus data

For the corpus study, we create richly annotated datasets from naturalistic language production in four broad registers at the intersection of formality and mode (see Section 4). That is, we contrast spoken and written language as well as formal and informal situations (see Figure 1). This selection is not intended to constitute registers at end points of the formality spectrum, but rather salient and commonly recognized text categories along this continuum. Each of the selected registers is internally heterogeneous with regard to the associated communicative purposes while production circumstances of the situational context are largely stable within those registers.

Figure 1. Selection of four broad registers at the intersection between mode and formality.



We focus on one variety of English, namely British English. In particular, we compare language use in informal conversation between family members and friends in the Spoken BNC2014 (Love et al. 2017) with use in House of Commons debates from 2007 to 2014 (provided by the Political Mashup project; Marx & Schuth 2010). The written informal register is represented by the British English blogs part of the GloWbE-corpus (Davies 2013; Davies & Fuchs 2015) and the written formal register is derived from articles of the newspaper *The Independent*, which were published between 2016 and 2019 (JSI Newsfeed corpus, Bušta et al. 2017). Using these corpora, we investigate two grammatical alternations: the dative alternation as in (1) and the

future marker alternation as in (2). Below, we report the first case study, namely on the dative alternation with *give*.

(1) Dative alternation:

a. ditransitive dative: *Tom gave Mary the book*

b. prepositional dative: *Tom gave the book to Mary*

(2) Future marker alternation:

a. *will*: *This will happen*

b. *be going to*: *This is going to happen*

Equivalent corpora were selected for the Dutch language, in which the corresponding grammatical alternations will be investigated. By examining two alternations, we aim to capture patterns that transcend a single alternation. This is also the motivation for examining these alternations in Dutch. Such a cross-variable, cross-linguistic comparison may elucidate register effects specific to one alternation, language, or both. This project serves as a pilot study for more research in a similar vein.

3.2.2 Experimental track

In the experimental track, we will conduct rating task experiments, in which we present participants with corpus sentences in context. Participants will read excerpts from two registers and rate the naturalness of equivalent variants on a continuous scale by means of a slider bar. The experimental

design of the project is inspired by previous work by Bresnan & Ford (2010) who developed a 100-split task and conducted an experiment on the dative alternation, presenting American and Australian English speakers with both alternatives in context (see also Ford & Bresnan 2013). Participants were asked to distribute 100 points between both variants based on their intuition of how likely the variants are given the context. That is, they were able to make graded judgments, for instance, they could distribute 30 points to one variant and 70 to the other. The authors found a positive correlation between the corpus probabilities and the experimental ratings, suggesting that speakers are sensitive to the probabilistic grammar of their variety. Klavan & Divjak (2016) provides an overview of similar work that compares corpus models and human rating performance and a discussion as to which extent the cognitive "reality" of probabilistic linguistic knowledge is captured by statistical models of corpus data and whether supplementary rating experiments can shed light on this question. The four studies that Klavan & Divjak review show that corpus and experimental data converge with regard to probabilistic grammatical choices.

4. Case study: The dative alternation in English

To illustrate a corpus-based variationist approach to study register, we will present a case study on the dative alternation with *give* in English (see Szmrecsanyi 2019, for a general introduction to different steps taken in variationist research practice). The dative alternation is one of the most extensively studied alternations in English (e.g., Bresnan et al. 2007; Röthlisberger et al. 2017; Wolk et al. 2013). The two constituents, which take the role of the recipient and the theme, alter in position, as shown in (3):

(3) a. ditransitive construction: *he's got to give* [*me*]$_{\text{recipient}}$ [*the money*]$_{\text{theme}}$ (BNC2014 SU82, S0041)

b. prepositional dative: *they don't give* [*pilot's licenses*]$_{\text{theme}}$ *to* [*idiots*]$_{\text{recipient}}$ (BNC2014 S3JF, S0227)

The language-internal factors contributing to the choice of the variant are well-understood; for instance, pronominality of the constituents, length and syntactic complexity of the constituents, or animacy of the recipient have been shown to condition the choice of the variant (for a detailed literature review see Gerwin 2014: Ch. 2; Röthlisberger 2018a: Ch. 2; see also Zehentner 2019: Ch. 3, for a historical account). Given that these factors have been shown to be influenced in their relative importance by variety as language-external factor (see Bresnan & Hay 2008; Röthlisberger et al. 2017; Szmrecsanyi et al. 2017), examining the dative alternation is particularly germane to investigating the variability of probabilistic constraints across registers (RQ1 and RQ2).

4.1 Variable context

In the four selected corpora, tokens of the verb *give* with its surrounding context were extracted semi-automatically.[2] First, all verb forms of *give* that were tagged as verbs were automatically matched and the utterances and sentences in which they occurred were extracted from the corpora.[3] Then, the variable context was coded manually.

Only tokens that can occur in both the ditransitive construction and the prepositional dative were considered for inclusion in the dataset.[4] Thus, in a first step, all invariable, incomplete tokens (i.e., when one constituent was missing), or mistagged tokens (e.g., *given* used as a preposition, conjunction, or adjective, but tagged as a verb) were filtered out. Following previous literature (e.g. Röthlisberger et al. 2017: 679; Theijssen et al. 2013: 232; Wolk et al. 2013: 389-391), invariable tokens would be, for example, formulaic and fixed expressions as in (4a), tokens in which *to* depends on the theme (4b),

[2] *Give* was chosen as a test case since it is a prototypical ditransitive dative verb based on its frequency (Bresnan & Hay 2008: 248; Gries 2003: 24; Gerwin 2014: 34, 107).
[3] The House of Commons corpus was available as raw text only. For the purpose of data extraction, a tagged version was created with the *spacy* (version 2.0) POS-tagger for English in Python 3. For all other corpora, the versions tagged by the corpus compilers were used. In relying on the automatically POS-tagged versions, mis-tagged instances of *give* were missed out.
[4] The percentage of tokens retained after the filtering process varies across corpora: 34.4% of the hits were identified as variable in written formal, 32.8% in spoken formal, 48.8% in written informal, and 57.6% in spoken informal registers respectively.

constructions with particle verbs as in (4c), or passivized constructions (4d).[5] Furthermore, constructions occurring in relative clauses were excluded when one constituent was realized in the independent clause (4e). While it would be possible to use the prepositional dative in (4e), i.e. *one of the best pieces of advice I can give to you*, sentences with a relativized constituent in the ditransitive construction do not have a canonical word order for ditranstive datives (cf. Wolk et al. 2013: 391).

(4) a. *I am happy to give way to the honourable Gentleman* (House of Commons, uk.proc.d.2007-01-22, uk.m.10652)

b. *The silver card will give access to unlimited off-peak films for 10 a month* (GloWbE-GB, blogs)

c. *oh I forgot to give it back to her* (Spoken BNC2014, S4PF, S0325)

d. *Mourinho was given a four-year contract extension* (The Independent, 02/01/2016)

e. *This is perhaps one of the best pieces of advice I can give you* (GloWbE-GB, blogs)

[5] Whenever it was unclear whether a construction can occur in both ditransitive or prepositional dative, a query in the iWeb corpus (Davies, 2018-) was carried out. When this query resulted in at least 10 instances of both variant forms, the token was retained, e.g., *give birth to*, for which also tokens in the ditransitive variant were found, for example: *Neither of them can wish the child did not exist, or regret the young girl's decision to give it birth.* (https://www.lrb.co.uk/the-paper/v36/n07/thomas-nagel/an-invitation-to-hand-wringing). This procedure shows that even though linguistic intuitions suggest otherwise and despite a strong bias of some lexemes to one or the other variant, both variants can be found in language use (Bresnan et al. 2007: 75).

After filtering for the variable context, we extracted, for each register, 650 randomly chosen tokens, half of which occurred in the ditransitive construction and the other half in the prepositional dative.[6] Overall, this resulted in a dataset of 2,600 tokens which were annotated for a range of language-internal constraints that play a role in the choice-making process between the two dative variants as previous research on the dative alternation has shown (summarized in Table 1).[7] Note that we opted for such a balanced dataset since we are not primarily interested in identifying differences in the frequencies of dative variants across registers, but in *why* dative variants should be more (or less) frequent in those registers.

Table 1. Overview of all constraints, their levels, and predictions for the ditransitive dative.[8]

Factor	**Levels**	**Predictions according to the literature**	**Literature**
Constituent length (WEIGHTRATIO =	continuous, length in characters	short recipients and long themes	Bresnan et al. (2007); Bresnan

[6] Since the parliamentary proceedings are not verbatim transcriptions of what has been said in political debates, the random subset was verified manually with regard to the actual language use by listening to the debate recordings available on https://www.parliamentlive.tv/Commons. In total, 148 tokens were filtered out for the following reasons: the speaker did not originally use *give*, the speaker used a benefactive construction with *for*, one of the constituents was missing or the constituents occurred in a non-canonical word order. Occasionally, the utterance could not be found or the recording was not available (9 tokens). In some tokens, the other variant was used originally (17 times the prepositional dative instead of the ditransitive dative, 3 times the ditransitive dative instead of the prepositional dative).

[7] Note that we largely followed annotation guidelines described in Röthlisberger (2018b).

[8] The table refers to effects found in logistic regression analyses unless indicated otherwise.

RECIPIENT/THEME)		favor the ditransitive dative	& Hay (2008); Theijssen et al. (2013); Röthlisberger et al. (2017)
RECIPIENT/THEME PRONOMINALITY	pronominal vs. non-pronominal	pronominal recipients and non-pronominal themes favor the ditransitive dative	Bresnan et al. (2007); Bresnan & Hay (2008); Theijssen et al. (2013); Röthlisberger et al. (2017)
RECIPIENT/THEME COMPLEXITY	simple vs. complex	simple recipients and complex themes favor the ditransitive dative	Röthlisberger et al. (2017); Röthlisberger (2018)
RECIPIENT/THEME FREQUENCY	continuous; register-specific normalized frequencies of the constituent head	high-frequent recipients and low-frequent themes favor the ditransitive dative	distribution in Röthlisberger (2018a)
RECIPIENT/THEME DEFINITENESS	definite vs. indefinite	definite recipients and indefinite themes favor the ditransitive dative	Bresnan et al. (2007); Theijssen et al. (2013); Röthlisberger et al. (2017)
RECIPIENT/THEME ANIMACY	animate vs. inanimate	animate recipients and inanimate themes favor the ditransitive dative	RECIPIENTANIMACY: Bresnan et al. (2007); Bresnan & Hay (2008); Theijssen et al. (2013); Röthlisberger et al. (2017) THEMEANIMACY: Röthlisberger (2018a)
VERBSENSE	abstract, communication, transfer	abstract and transfer verb senses favor the ditransitive dative	distribution in Röthlisberger (2018a), but see coefficients in Bresnan & Hay (2008)

REGISTER	spoken informal (conversations), spoken formal (parliamentary debates), written informal (blogs), written formal (newspaper)	spoken informal favors the ditransitive dative, higher probability of prepositional dative in written formal texts	distribution of dative variants, e.g. Bresnan et al. (2007) & Röthlisberger (2018a)

4.2 Language-internal constraints

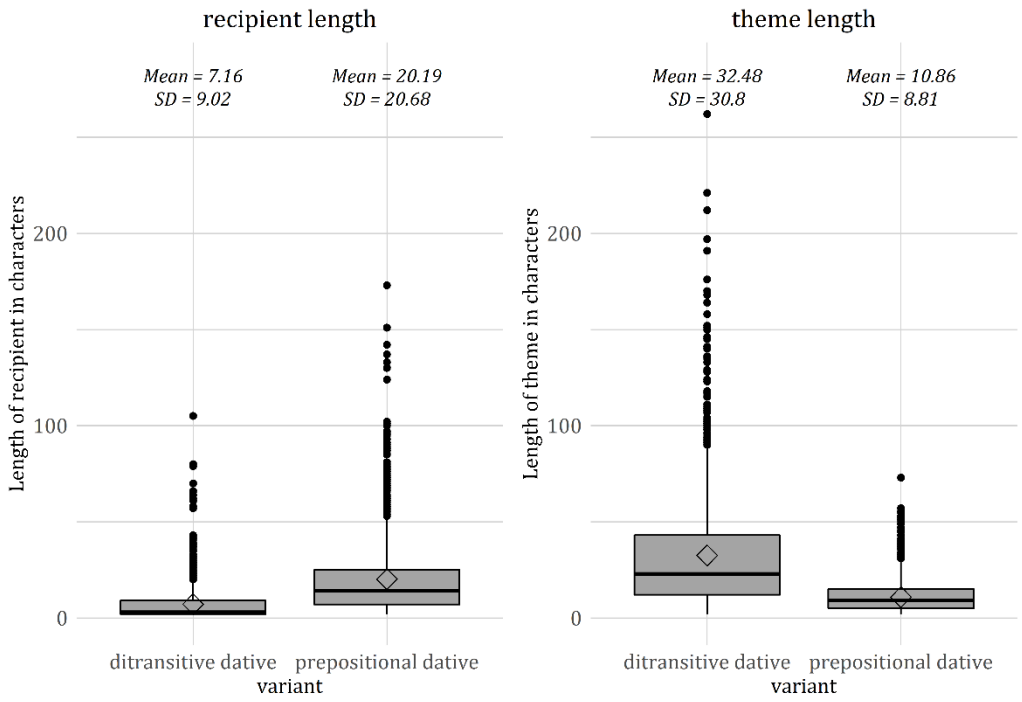
4.2.1 Constituent length

The dative alternation is subject to the end weight principle (Behaghel 1909), according to which short elements tend to precede longer elements (see also Hawkins 1994). For the dative alternation, this means that the ditransitive dative construction is preferred when the recipient constituent is shorter than the theme and the prepositional dative is preferred when the theme is shorter than the recipient (e.g., Bresnan et al. 2007; Bresnan & Hay 2008).

For the present analysis, we calculated the length of the constituent as the number of characters (including whitespaces). The general tendency for shorter constituents to occur in the first slot in the dative alternation is also reflected in the distribution of the constituent length in our dataset. As Figure 2 shows, recipients are longer in prepositional dative tokens compared to ditransitive dative tokens (unpaired t-test: $t = -20.8$, $df = 1776.2$, $p < 0.001$, Cohen's $d = 0.82$), and themes are longer in ditransitive tokens compared to

prepositional dative tokens (unpaired t-test: $t = 24.3$, $df = 1510.4$, $p < 0.001$, Cohen's $d = 0.96$).

Figure 2. Boxplots for constituent length. Squares indicate the mean, the thick lines inside the boxes indicate the median.[9]



We combined the length measures of the constituents into a single measure, the weight ratio, by dividing the length of the recipient by the length of the theme. As shown by Shih & Grafmiller (2011), length operationalizations are highly correlated, and model comparisons between models including ratio

[9] Figures were created in R (version 3.6.2) using *ggplot2*.

or separate length predictors for each constituent yielded a better model fit for the model with the ratio as length predictor. Thus, the ratio was calculated by dividing recipient length by theme length. As for the token in (5), the recipient is 6 characters long and the theme is 42 characters long, resulting in a weight ratio of 6/42 = 0.143). This weight ratio was further log-transformed and standardized for the logistic regression analysis (as described in the Analysis section).

(5) *Mr Grieve said it was important to give* [*people*]$_{\text{recipient}}$ [*the chance to change their minds on Brexit*]$_{\text{theme}}$ (The Independent, 03/02/2018)

4.2.2 Pronominality

Previous research has shown that the ditransitive construction is more likely when the recipient is pronominal and when the theme is non-pronominal (e.g., Bresnan et al. 2007). Pronominal themes, on the other hand, favor the prepositional dative. Constituents were coded as pronominal when their head was a pronoun (either definite/personal, indefinite, or demonstrative; see in 6a) and as non-pronominal when their head was a (proper) noun (see in 6b).

(6) a. *so you can give* [*one*] $_{\text{pronominal}}$ *to* [*somebody else*] $_{\text{pronominal}}$ *so I'm a Cylon yay* (Spoken BNC2014, SAUR, S0192)

b. *New labelling will give* [*consumers*]$_{\text{non-pronominal}}$ [*more power to say no to the palm oil that fuels deforestation*]$_{\text{non-pronominal}}$ (GloWbE-GB, blogs)

In our dataset, there are more pronominal recipients in the ditransitive dative than in the prepositional dative (see Table 2). This stands in contrast to the distribution of pronominal themes, in that there are more pronominal themes in the prepositional dative compared to the ditransitive dative (see Table 3).

Table 2. Cross-table RECIPIENTPRONOMINALITY (χ^2 =516.3, df = 1, p < .001, φ = 0.45).

	ditransitive dative	prepositional dative	row total
pronominal	819 (63%)	249 (19.2%)	1068
non-pronominal	481 (37%)	1051 (80.8%)	1532
column total	1300	1300	2600

Table 3. Cross-table THEMEPRONOMINALITY (χ^2 =228, df = 1, p < .001, φ = 0.3).

	ditransitive dative	prepositional dative	row total
pronominal	27 (2.1%)	273 (21%)	300
non-pronominal	1273 (97.9%)	1027 (79%)	2300
column total	1300	1300	2600

4.2.3 Complexity

As shown in Röthlisberger et al. (2017), the likelihood for the prepositional dative increased when the recipient is syntactically complex, i.e., when the

constituent is postmodified (see 7a). The ditransitive dative is favored when the theme is complex (see 7b). This effect is in accordance with the 'Easy First' bias, according to which elements that are easier to retrieve from memory are placed first in an utterance (MacDonald 2013; see also discussion in Section 5).

(7) a. *Raandom will give* [*credit*]$_{\text{simple}}$ *to* [*jokes from other sites*]$_{\text{complex}}$ (GloWbE-GB, blogs)

b. *The segment gave* [*viewers*]$_{\text{simple}}$ [*the first lesbian wedding shown on network TV*]$_{\text{complex}}$ (The Independent, 11/01/2019)

In the present study, we distinguish simple constituents (without postmodification) from complex constituents that include restrictive postmodifications, adding new information for the identification of the constituent's head. Postmodifications were either relative clauses, appositions, *to*-/*that*-complement clauses, prepositional phrases, or coordinated constituents, as well as abbreviations in brackets (*the Special Interest Group (SIG)*) or adverbs following the head (*anyone else*). The distribution across variants shows that there are more complex recipients in the prepositional dative compared to the ditransitive dative variant (see Table 4) and more complex themes in the ditransitive dative than in the prepositional dative (see Table 5).

Table 4. Cross-table RECIPIENTCOMPLEXITY. ($\chi^2 = 249.2$, $df = 1$, $p < .001$, $\varphi = 0.31$).

	ditransitive dative	prepositional dative	row total
simple	1248 (96%)	960 (74%)	2208
complex	52 (4%)	340 (26%)	392
column total	1300	1300	2600

Table 5. Cross-table THEMECOMPLEXITY ($\chi^2 = 736.1$, $df = 1$, $p < .001$, $\varphi = 0.53$).

	ditransitive dative	prepositional dative	row total
simple	611 (47%)	1238 (95.2%)	1849
complex	689 (53%)	62 (4.8%)	751
column total	1300	1300	2600

The complexity predictor is also related to the principle of end weight, and often length measures and syntactic complexity go hand in hand: a complex constituent tends to be longer than a simple constituent, but simple constituents, as operationalized here, can also include pre-modifications to the syntactic head (and thus be of comparable length to complex constituents). Wasow & Arnold (2003) collected acceptability judgments of dative sentences containing constituents with the same length but varying complexity. Their results show that complexity and length measures have distinct effects and that including both best captures the variation of constituent ordering as opposed to only including one or the other.

4.2.4 Frequency

As reliably shown in psycholinguistic research, high-frequent words are more accessible than infrequent words, evidenced by studies on reading and reaction times (e.g., Rayner & Duffy 1986; Balota & Chumbley 1984). Following the assumption of an Easy First bias (MacDonald 2013), we expect that the ditransitive dative is more likely when the recipient is highly frequent and that the prepositional dative is more likely when the theme is highly frequent.

To account for such potential frequency-related effects, relative normalized frequencies for the head lemmata were included in the analysis. Frequency counts were corpus-specific, as opposed to a uniform count across registers, since frequency varies dependent on register and it might be the case that this affects language users differently. For example, the word *government* has a normalized frequency of 4378 occurrences per 1 million words in the parliamentary debates corpus, while it has only a normalized frequency of 673 occurrences per million in the newspaper corpus and even lower frequencies in the informal corpora with 52 occurrences per million words in the SpokenBNC2014 and 428 occurrences per million words in the British English blogs section of the GloWbE. Note that frequency is not commonly included as a language-internal constraint in research on the dative alternation (but see Röthlisberger 2018a). The distributions of normalized frequency counts in

Figure 3 show that, in our dataset, recipients are more frequent in the ditransitive dative than in the prepositional dative (unpaired *t*-test: $t = 11.1$, $df = 2575$, $p < .001$, Cohen's $d = 0.44$) and themes are more frequent in the prepositional dative compared to the ditransitive dative (unpaired *t*-test: $t = -15.8$, $df = 1364.8$, $p < .001$, Cohen's $d = 0.62$).

Figure 3. Boxplots for constituent frequency (pmw). Square symbols indicate the mean, the thick lines inside the boxes indicate the median.

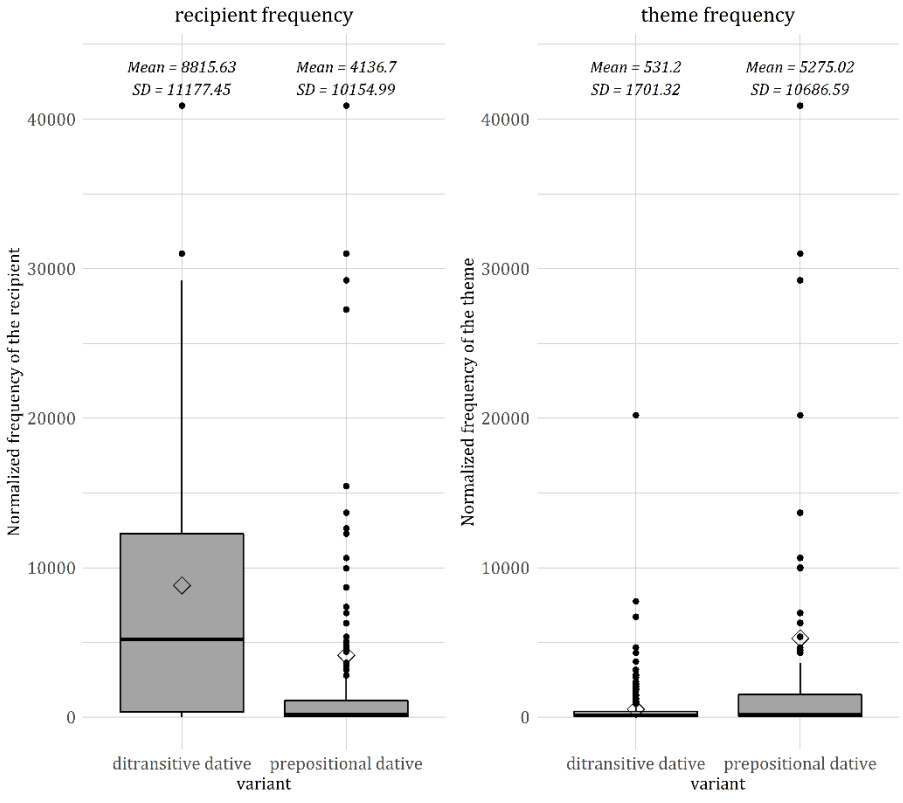


Figure 3 also shows that there are some very high-frequency constituents. These high-frequency constituents constitute pronouns such as *me*, *you*, *it*, *them* and demonstrative pronoun *that*. The fact that the mean lies outside of the interquartile range (indicated by the boxes) shows that the frequency distributions are skewed. Therefore, these predictors were logarithmized before entering them into the regression analysis (see Section 4.3).

4.2.5 Definiteness

Definiteness is related to information status of the constituent, which affects the ease of processing. Indefinite referents are less accessible because they generally refer to new information (Chafe 1976). Previous studies have found that the odds for the ditransitive dative increase when the recipient is definite and the theme is indefinite (e.g., Bresnan et al. 2007; Theijssen et al. 2013).

We distinguish between definite versus indefinite constituents following the coding scheme by Garretson et al. (2004; see also Röthlisberger 2018a: 66-67). Definite constituents have a definite pronoun, a proper noun, or a noun preceded by a definite determiner as a head (see in 8a). Indefinite constituents are headed by indefinite pronouns or nouns preceded by an indefinite determiner as well as bare nouns (see in 8b).

(8) a. *we give* [*our full support*]$_{\text{definite}}$ *to* [*this proposal*]$_{\text{definite}}$ (House of Commons, uk.proc.d.2010-01-05, uk.m.10659)

b. *The kitchen gives* [*audio instructions in French*]$_{\text{indefinite}}$ *to* [*cooks who are learning that language*]$_{\text{indefinite}}$ (GloWbE-GB, blogs)

In the present dataset, more definite recipients are ditransitive datives compared to prepositional datives (see Table 6) and more definite themes occur in prepositional dative variants compared to ditransitive dative (see Table 7).

Table 6. Cross-table RECIPIENTDEFINITENESS ($\chi^2 = 159.1$, $df = 1$, $p < .001$, $\varphi = 0.25$).

	ditransitive dative	prepositional dative	row total
definite	1138 (87.5%)	868 (66.8%)	2006
indefinite	162 (12.5%)	432 (33.2%)	594
column total	1300	1300	2600

Table 7. Cross-table THEMEDEFINITENESS ($\chi^2 = 32$, $df = 1$, $p < .001$, $\varphi = 0.11$).

	ditransitive dative	prepositional dative	row total
definite	355 (27.5%)	490 (37.7%)	845
indefinite	945 (72.5%)	810 (62.3%)	1755
column total	1300	1300	2600

4.2.6 Animacy

Based on previous findings, we assume that the ditransitive construction is more likely when the dative construction includes an animate recipient (Bresnan et al. 2007, see sample sentences in 9).

(9) a. *okay well I 'll give* [*you*]$_{\text{animate}}$ [*that twenty quid*]$_{\text{inanimate}}$ *then* (Spoken BNC2014 SE88, S0083)

b. *They could have given* [*borrowing powers*]$_{\text{inanimate}}$ *to* [*Scotland and Wales*]$_{\text{inanimate}}$, *but that hasn't happened*. (House of Commons, uk.proc.d.2009-03-17, uk.m.14137)

The annotation scheme for this study followed Wolk et al. (2013) who defined five categories (i.e., 'human', 'animate', 'collective', 'locative', 'temporal') based on the guidelines for animacy by Zaenen et al. (2004). This initial annotation scheme with five distinctions was reduced to a binary one (animate vs. inanimate), in which humans, animals, and human- or animal-like entities (e.g., characters of video games) were coded as animate, and collective, locative, and temporal constituent head nouns as inanimate. This decision for a binary predictor serves to make the model simpler as it reduces the number of coefficients and potential causes for multicollinearity such as data sparseness.

As shown in Table 8, more inanimate recipients are found in the prepositional dative compared to the ditransitive dative in the present data. Table 9 shows that there are more animate themes in the prepositional dative

than in the ditransitive dative. This finding is in line with previous research (Röthlisberger 2018a). In addition, there is a highly skewed distribution for theme head nouns across registers since the subset of the spoken formal register did not include any animate themes.

Table 8. Cross-table RECIPIENTANIMACY. ($\chi^2 = 122.8$, $df = 1$, $p < .001$, $\varphi = 0.22$).

	ditransitive dative	prepositional dative	row total
animate	983 (75.6%)	714 (54.9%)	1697
inanimate	317 (24.4%)	586 (45.1%)	903
column total	1300	1300	2600

Table 9. Cross-table THEMEANIMACY($\chi^2 = 6.8$, $df = 1$, $p = .016$, $\varphi = 0.05$).

	ditransitive dative	prepositional dative	row total
animate	6 (0.5%)	19 (1.5%)	25
inanimate	1294 (99.5%)	1281 (98.5%)	2575
column total	1300	1300	2600

4.2.7 Verb sense

Previous investigations found that communicative use and transfer use of the verb *give* increase the odds for the prepositional dative compared to abstract use of *give* in American and New Zealand English varieties (Bresnan & Hay 2008).

To account for semantic nuances of the dative expression, verb sense was included in the dataset. Depending on the semantics of the theme lemma, the verb *give* can be categorized into three different meanings (cf. Bresnan et al. 2007: 78, 85; Bresnan & Hay 2008: 250): 'transfer' for physical transfer of a concrete object (10a), 'communication' for transfer of information (10b), or 'abstract' for transfer of a non-concrete object (10c). The categories 'abstract' and 'communication' can be distinguished in that the communication sense denotes exclusively communication events. Ambiguous cases were annotated as 'abstract'.

(10) a. transfer: *I gave the apples to my dad to make jam/jelly*. (GloWbE-GB Blogs)

b. communication: *Could I give him an exact example of this?* (House of Commons, uk.proc.d.2008-10-15, uk.m.11817)

c. abstract: *Israeli Prime Minister Benjamin Netanyahu gave his support to the proposals on Sunday* (The Independent, 18/11/2016)

Table 10 shows that there are more ditransitive dative tokens with an abstract verb sense compared to prepositional datives. Furthermore, we find more tokens with a verb sense of transfer in the prepositional dative than in the ditransitive dative.

Table 10. Cross-table VERBSENSE ($\chi^2 = 72.5$, $df = 2$, $p < .001$, $V = 0.17$).

	ditransitive dative	prepositional dative	row total
abstract	963 (74.1%)	782 (60.2%)	1745
communication	151 (11.6%)	162 (12.5%)	313
transfer	186 (14.3%)	356 (27.4%)	542
column total	1300	1300	2600

4.3 Analysis

A multi-level mixed effects regression model was fitted in R (version 3.6.2, R Core Team 2019) with the package *lme4* (Bates et al. 2015) for logistic mixed effects regression models (family "binomial"). The model predicts the odds for the prepositional dative. Treatment coding was used and the reference levels of categorical predictors were set to those levels that are the most common ones for the ditransitive dative construction (cf. Table 1). The reference level for REGISTER was set to "spokeninformal". All continuous predictors were logarithmized, centered and standardized, in order to not violate basic assumptions of logistic regression models. To account for lexical or speaker-related idiosyncrasies, the head lemma of the themes and the recipients as well as speaker identity were entered as random effects.

We attempted to enter all possible interactions with REGISTER into the model, but this yielded high multicollinearity. This statistical problem arises when predictors are not independent of each other and share the same part of the variance in regression analyses. As a result, it is impossible to tease apart the relative importance of the correlated predictors (York 2012). In our case, multicollinearity arises due to the fact that predictors like pronominality, definiteness, length and frequency are correlated because, for example, pronominal constituents are definite, short and highly frequent. In addition, the levels of the predictors are not uniformly distributed across registers and dative variants. For example, there are few pronominal themes in the spoken formal and written formal parts of the dataset, and (as for the overall tendency) in both formal registers, pronominal themes occur mostly or even exclusively in prepositional dative variants. In order to be able to interpret the model, we opted to simplify the model structure with regard to interaction terms. Alternatively, we considered the option of residualizing the collinear predictor(s) against REGISTER as a remedy since the collinearity only arises when predictors are put in interaction with REGISTER. However, this practice is dispreferred as it biases the residualizer (i.e., REGISTER in the present analysis), meaning that the shared variance will be attributed exclusively to the residualizer, which would not lead to a better interpretability of the effect of individual predictors (cf. Wurm & Fisicaro 2014; see also York 2012).

Therefore, we chose those language-internal variables that appeared to be stable in interaction with REGISTER (i.e., RECIPIENTDEFINITENESS and THEMEDEFINITENESS) and theoretically interesting (WEIGHTRATIO; cf. Bresnan & Ford 2010) to be included in two-way interaction terms with REGISTER.[10] All language-internal predictors were entered as main effects. The model selection process followed recommendations by Gries (2015; see also Zuur et al. 2009, Ch. 5), i.e., we applied a backward selection process and performed model comparisons (using the *anova()*-function) to find the best fit. First, the random effect structure had been simplified by removing those random effects that did not significantly improve the explanatory power of the model. Then, non-significant interactions were subsequently removed, followed by non-significant main effects.

The final model has the structure shown in (11).

(11) VARIANT ~ (1 | THEMEHEAD) + RECIPIENTANIMACY + THEMECOMPLEXITY + RECIPIENTDEFINITENESS + THEMEDEFINITENESS + RECIPIENTPRONOMINALITY + THEMEPRONOMINALITY + VERBSENSE +

[10] A reviewer suggested to fit a model only with the interaction between REGISTER and WEIGHTRATIO. When we do so, we find a significant interaction between WEIGHTRATIO and REGISTER (β = -0.9785, p = 0.045), suggesting that the effect of WEIGHTRATIO is stronger in the spoken informal register compared to the written formal register. Thus, when the weight ratio becomes larger (and the recipient longer than the theme), the prepositional dative becomes more likely in the spoken informal register compared to the written formal register. This model, however, does not converge and has a poorer fit to the data than the model reported here.

WEIGHTRATIO + REGISTER + REGISTER*RECIPIENTDEFINITENESS +

REGISTER*THEMEDEFINITENESS

Model evaluations show that the concordance index *C* is high with 0.974, indicating that the model is able to discriminate between the ditransitive and the prepositional dative constructions. The model mispredicts 220 tokens out of the dataset of 2,600 tokens, which results in an overall accuracy of 91.5% (baseline = 50%). Pseudo-R^2 (following Nakagawa & Schielzeth 2013) returns a conditional pseudo-R^2 of 0.846, indicating that roughly 85% of the variance in the data is explained by the overall model, and a marginal pseudo-R^2 of 0.699, indicating that 69.9% of the variance in the data is explained by the fixed effects alone. To test for multicollinearity, the condition index was calculated with the intercept included following Belsley et al. (1980). The result $\kappa = 13.45$ indicates that there is medium collinearity. Note, however, that only condition indices above 30 indicate "potentially harmful collinearity" (Baayen 2008: 282).

4.4 Results

As shown in Table 11, there is only one random effect that improves the model fit, namely the one for theme head lemma, which adjusts the intercept according to lexical idiosyncrasies of the theme, similar to previously reported lexical

effects in the dative alternation (Bresnan & Ford 2010; Röthlisberger et al. 2017). The largest adjustments to the intercept are found for *evidence*, *birth*, *it*, *they*, and *light* favoring the prepositional dative and *go*, *say*, *lead*, *assurance*, and *chance* favoring the ditransitive dative construction. The random effects for speaker and recipient head lemma did not significantly contribute to explaining the variance. The absence of a random effect for speaker could be due to missing author information in the written registers and the resulting coding of speaker identity equal to text identity. Given the random effect structure of our final model, the theme seems to be lexically more bound to one of the dative variants than the recipient.

Table 11. Model output.[11] Predictions are for the prepositional dative.

Predictor	***Coefficient***	***Odds Ratio***	***SE***	***p***
(Intercept)	-5.22	0.01	0.49	**<0.001**
RECIPIENTANIMACY				
animate ⇒ inanimate	0.892	2.44	0.18	**<0.001**
THEMECOMPLEXITY				
complex ⇒ simple	2.135	8.45	0.23	**<0.001**
RECIPIENTDEFINITENESS				
definite ⇒ indefinite	2.204	9.06	0.59	**<0.001**
THEMEDEFINITENESS				
indefinite ⇒ definite	1.219	3.38	0.38	**0.001**
WEIGHTRATIO				
log(recipient/theme length)	1.879	6.55	0.16	**<0.001**
RECIPIENTPRONOMINALITY				

[11] For the sake of brevity, the reader is referred to Levshina (2015: 261-262) for further information on log-odds and odds ratios.

pronominal ⇒ non-pronominal	2.608	13.57	0.28	**<0.001**
THEMEPRONOMINALITY				
non-pronominal ⇒ pronominal	1.545	4.69	0.59	**0.009**
VERBSENSE				
abstract ⇒ communication	1.435	4.20	0.32	**<0.001**
abstract ⇒ transfer	0.761	2.14	0.29	**0.01**
REGISTER				
spoken informal ⇒ spoken formal	0.791	2.21	0.39	**0.042**
spoken informal ⇒ written informal	0.791	2.20	0.38	**0.035**
spoken informal ⇒ written formal	-0.613	0.54	0.39	0.107
REGISTER * THEMEDEFINITENESS				
spoken formal + definite	-1.879	0.17	0.67	**0.005**
written informal + definite	-1.187	0.63	0.72	0.539
written formal + definite	-0.416	0.45	0.68	0.101
REGISTER * RECIPIENTDEFINITENESS				
spoken formal + indefinite	-1.770	0.15	0.52	**0.006**
written informal + indefinite	-0.468	0.31	0.53	0.381
written formal + indefinite	-0.803	0.66	0.50	0.1100
Random Effect	σ^2			
ThemeHeadLemma	3.15			

Table 11 also shows the coefficients (i.e., log-odds) and odds ratios for the main effects and interactions found by the model. Negative coefficients (and odds ratios smaller than 1) disfavor the prepositional dative, while positive coefficients (and odds ratios larger than 1) favor the prepositional dative. The directions of the main effects are in line with previous findings. The largest effect is found for RECIPIENTPRONOMINALITY, with the odds for the prepositional dative increasing by 13.57 when the recipient is not a pronoun. The effect for THEMEPRONOMINALITY is smaller: when the theme is pronominal

the odds for the prepositional dative increase by 4.69. These effects for pronominality are in line with the effect found for WEIGHTRATIO: since pronouns are short and generally not postmodified, the shorter constituent of the two is placed first in the dative construction. In other words, the larger the weight ratio, the more likely the prepositional dative. Similarly, the prepositional dative becomes more likely when the recipient is indefinite and inanimate, while the theme is definite and simple. Dative constructions with communication or transfer as verb sense prefer the prepositional dative compared to constructions with abstract meaning. With regard to REGISTER, a significant main effect was found when spoken informal conversations were compared to written informal blog texts as well as to spoken formal parliamentary debates, indicating that the prepositional dative is more likely in written informal and spoken formal language than in conversations. The comparison between spoken informal and written formal registers does not reach significance. All effect sizes for REGISTER are smaller compared to the ones of the language-internal predictors. The main effects for RECIPIENTCOMPLEXITY, RECIPIENTHEADFREQUENCY, THEMEHEADFREQUENCY, and THEMEANIMACY were not significant.

Figure 4. Effect plot of interaction between REGISTER and RECIPIENT DEFINITENESS.

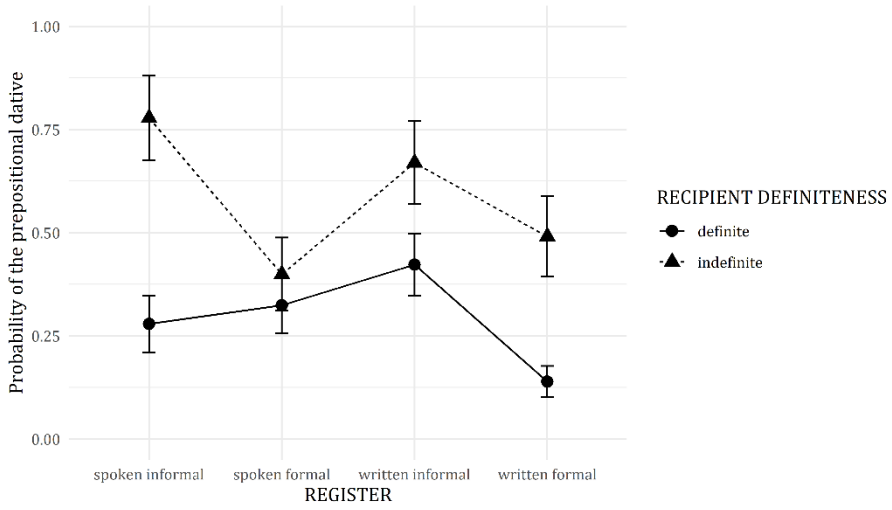
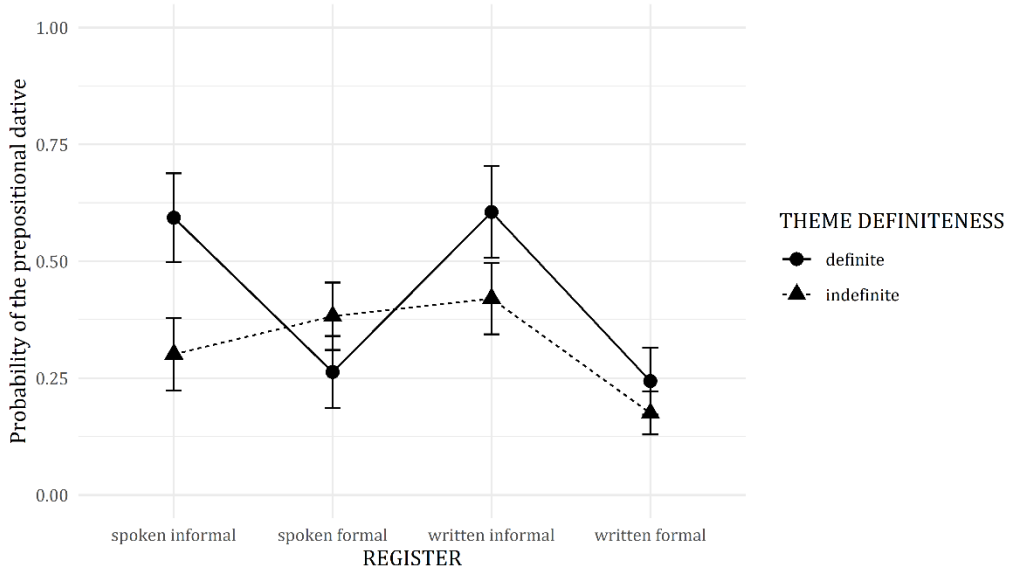


Figure 5. Effect plot of the interaction between REGISTER and THEME DEFINITENESS.



As for the interactions, we found an interaction between REGISTER and RECIPIENTDEFINITENESS (Figure 4) and one between REGISTER and

THEMEDEFINITENESS (Figure 5). There is no interaction between REGISTER and WEIGHTRATIO. Figure 4 and Figure 5 show the probability of the prepositional dative (y-axis) for both levels of the definiteness predictors across four registers (x-axis). The dot symbols (connected with a solid line) show the effect for definite constituents, the triangles (with a dashed line) show the effect for indefinite constituents. The wider the gap between the effect for definite and indefinite constituents, the stronger the effect for that predictor in the given register. Looking at the interaction between REGISTER and RECIPIENTDEFINITENESS (Figure 4), we find the strongest effect for recipient definiteness in the spoken informal register in that indefinite recipients strongly favor the prepositional dative. The smallest difference between definite and indefinite recipients is found in the spoken formal register. This difference is significant, whereas the comparison between the effect sizes in spoken informal and written informal or spoken informal and written formal registers are not. In addition, a significant interaction was found between REGISTER and THEMEDEFINITENESS (Figure 5), again in the spoken registers along the formality continuum. Interestingly, however, while all the other registers follow the general trend in that definite themes prefer the ditransitive construction, the direction of the effect of theme definiteness is reversed in the spoken formal parliamentary debates. Both formal registers show smaller effects for THEMEDEFINITENESS compared to the informal registers. To sum up,

additionally to the expected main effects, the model finds an interaction between REGISTER and RECIPIENTDEFINITENESS and an interaction between REGISTER and THEMEDEFINITENESS (see Table 12). These findings will be further discussed in the following section.

Table 12: Summary of the mixed effects regression results.

Predictor	**Result**
WEIGHTRATIO	the longer the relative length of the recipient, the more likely the prepositional dative (PD)
RECIPIENTPRONOMINALITY	PD more likely when recipient is not a pronoun
THEMEPRONOMINALITY	PD more likely when theme is a pronoun
RECIPIENTCOMPLEXITY	ns
THEMECOMPLEXITY	PD more likely when the theme is simple
RECIPIENTHEADFREQUENCY	ns
THEMEHEADFREQUENCY	ns
RECIPIENTDEFINITENESS	PD more likely when the recipient is indefinite
THEMEDEFINITENESS	PD more likely when the theme is definite
RECIPIENTANIMACY	PD more likely when the recipient is inanimate
THEMEANIMACY	ns
VERBSENSE	PD more likely when the verb sense is 'communication' or 'transfer'
REGISTER	PD more likely in spoken formal and written informal registers
REGISTER***** RECIPIENTDEFINITENESS	the effect of RECIPIENTDEFINITENESS is weaker in the spoken formal register

REGISTER***** THEMEDEFINITENESS	the effect of THEMEDEFINITENESS is reversed in the spoken formal register ⇒ PD more likely when the theme is indefinite

5. Discussion and conclusion

Linguistic variation across registers has so far widely been studied by employing Multidimensional Analysis (e.g., Biber 1988), a text-linguistic approach to register variation. The present contribution has outlined an approach to study register from a variationist perspective by combining corpus studies on two grammatical alternations with rating task experiments. This line of research is set apart from other research in corpus-based variationist linguistics in that it focuses on the interaction between language-internal constraints and register, rather than including register as a random effect or main effect only. Corpus results from a case study on the English dative alternation with *give* demonstrated how register modulates the probabilistic effects of definiteness of the constituents.

Analysis shows that the core grammar for the British English dative alternation is largely stable across the registers included in this study. Overall, the results are in line with the 'harmonic alignment effects' with syntactic

position (Bresnan & Ford 2010: 183f.; see also Bresnan et al. 2007: 80). That is, whichever constituent occurs in the first slot is likely to be discourse given, pronominal, animate, definite, and short or less syntactically complex. Importantly, such effects can be explained with the 'Easy First' bias in language production (MacDonald 2013). According to this bias, entities that are easier to retrieve from memory are placed earlier in utterances. Ease of retrieval has been evidenced for given, concrete, high-frequent, short, and less complex forms. Moreover, animate referents have been shown to be conceptually more accessible than inanimate ones (Branigan et al. 2008). While such biases are formulated for spoken language production, we assume that processing-related constraints also hold in writing since written text production similarly involves word retrieval and sentence planning (cf. Hayes & Flower 1980).

In contrast to previous assumptions in variationist linguistics, we do find interaction effects between register and language-internal constraints, specifically definiteness of the recipient and the theme. More precisely, the relative importance of recipient definiteness is modulated by the formality of the speech situation, contrasting informal conversations with parliamentary debates. In addition, we find an effect of formality in spoken language for theme definiteness. Interestingly, not only the magnitude of the theme definiteness effect is influenced but also the direction of the effect is reversed in the spoken formal situation compared to the spoken informal one. This suggests

that the probabilistic grammar of spoken registers varies as a function of formality.

As Figure 6 shows, the distribution is skewed with regard to recipient definiteness in spoken informal conversations, not only in terms of the distribution between definite and indefinite recipients overall, but also in terms of the distribution of dative variants, in that the ditransitive dative construction is clearly disfavored with indefinite recipients. In general, there are more indefinite recipients in the spoken formal register. Assuming that definite referents are more accessible than indefinite ones (cf. the Givenness Hierarchy by Gundel et al. 1993), this effect can be interpreted in relation to the Easy First principle. In spontaneous conversations, definite referents are placed first because they are easier to access and to process. Similarly to the distribution of recipient definiteness across registers, we also find more indefinite themes in the spoken formal register overall (Figure 7). It seems that, in general, more indefinite referents are used in parliamentary debates compared to informal conversations. This might be explained with the high frequency of definite pronouns in informal conversations, as opposed to the higher frequency of nouns in more informational registers (Biber 1988; Biber et al. 1999: 235f.).

Figure 6. Distribution of dative variants depending on RECIPIENT DEFINITENESS across four broad registers.

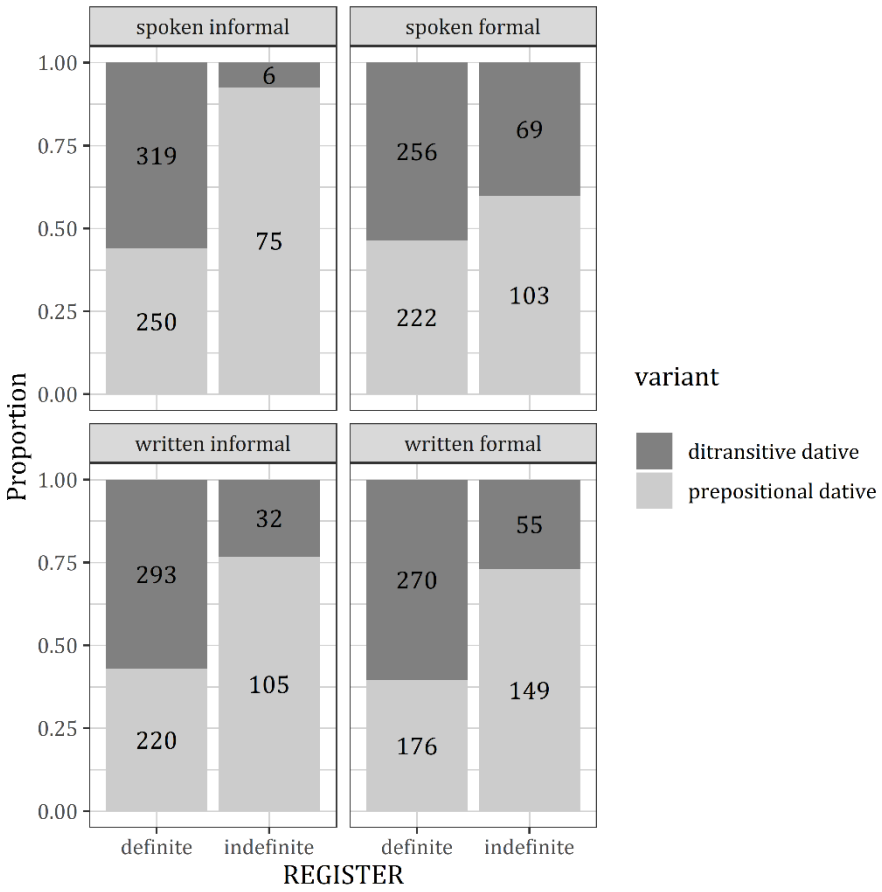
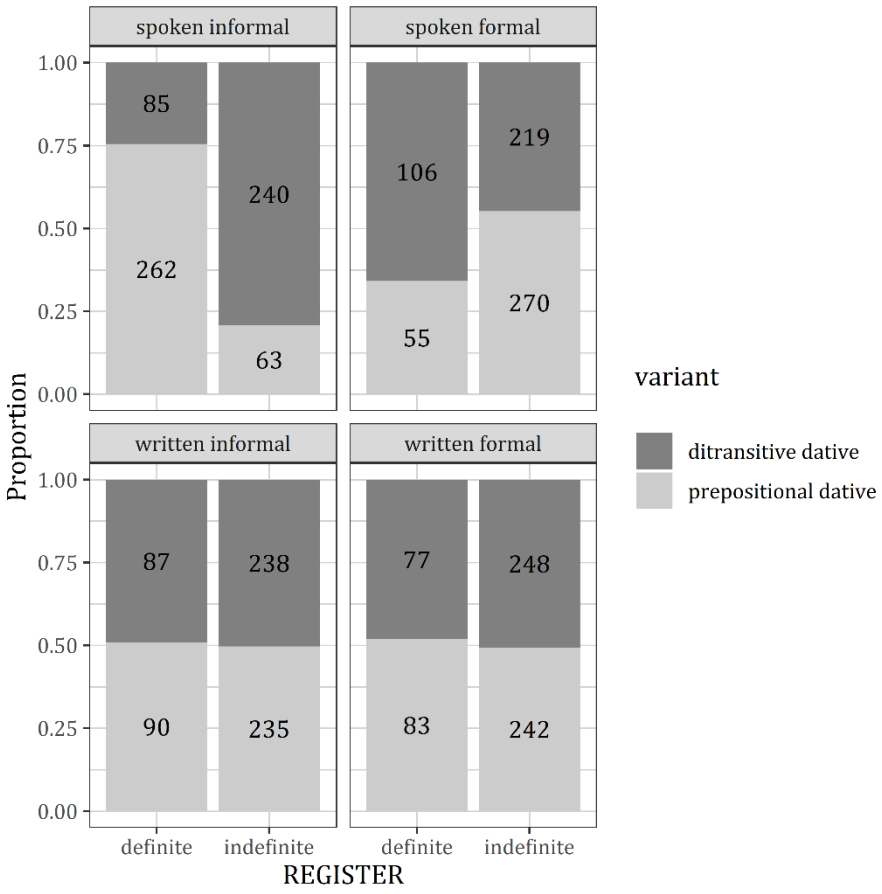


Figure 7. Distribution of dative variants depending on THEME DEFINITENESS across four broad registers.



There are several possible explanations for these interaction effects that pertain to (processing-related) characteristics of the situational context of the two spoken registers in the present study. One key difference is that parliamentary debates include pre-planned speeches which are essentially written-to-be-read. As such, some of the tokens in the dataset represent conceptually written language (Koch & Oesterreicher 2012) although the mode is spoken. Utterance planning in spontaneous conversations requires different

(real-time) cognitive capacities compared to conceptual written language production, where editing is possible (Biber 1988: 107). Another difference lies in the content of formal versus informal registers: the former is more informational or argumentative while the latter is often more personal and involved. Biber (1988: 107) argues that the distribution of features along this dimension in his MDA of English (Dimension 1) is related to production circumstances of speech. Thus, online processing-related principles (like the Easy First bias) should have a greater effect in spontaneous speech than in writing. Surprisingly, however, the effects for definiteness and for weight ratio do not differ between spoken informal and written formal registers, where we expected to find the largest differences caused by differential processing constraints in utterance planning. This could be due to the fact that the written registers included in the study here are rather heterogeneous in that they include several subgenres. That is, newspaper articles include not only news or sports report, but also articles from the sections 'lifestyle' or 'voices' with a more personal style. Similarly, the blogs that are part of the GloWbE corpus include texts with diverse communicative purposes (cf. Biber & Egbert 2018). In this sense, the selection of the written corpora for the study entails drawbacks and it might be due to this heterogeneity in the written registers that we do not find significant differences in probabilistic effects between spoken and written

registers. An MDA with the texts from which we drew our random sample could provide further insights into this issue.

On the other hand, the dative alternation is an alternation in which stylistic differences between registers are not obvious. Yet, the present analysis showed that we can also find interactions between register and language-internal constraints for the dative alternation. Future research will show whether the register-specific probabilistic grammars can be found reliably across different alternations (e.g., Grafmiller 2014; Jankowski 2013) and languages. If that is the case, we can – according to Guy's Grammatical Difference Hypothesis –contend that speakers of a language with multiple registers are in fact multilingual, having several probabilistic grammars according to which they make linguistic choices in different registers:

> When the CONTEXTS of use differ, different grammars are involved. […] [In] complicated situations with multiple grammars competing in a community, individuals may differ substantively in the contexts of variation; however, using different constraint effects stylistically will be equivalent to diglossic or bilingual behavior, rather than simple stylizing within one language (Guy 2015: 14f.; emphasis in original).

To conclude, our case study has confirmed that a shift towards investigating alternations across registers and genres other than the vernacular

seems warranted for variationist linguistics as "speech is not always the key locus of this kind of variation" (D'Arcy & Tagliamonte 2015: 279). We believe that variationist research can greatly benefit from a multi-feature design as well as a twofold methodology with both corpus study and experimental track. While we reported only a corpus study on one alternation here, we aim to apply such a multifaceted methodology in the larger project. By complementing observational data with experimental data, we will be able to make stronger claims about the cognitive system underlying probabilistic choice-making than with a single-method approach.

References

Arppe, Antti, Gilquin, Gaëtanelle, Glynn, Dylan, Hilpert, Martin & Zeschel, Arne. 2010. Cognitive corpus linguistics: Five points of debate on current theory and methodology. *Corpora* 5(1): 1–27. https://doi.org/10.3366/cor.2010.0001.

Baayen, R. Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Balota, David A. & Chumbley, James I. 1984. Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception & Performance* 10(3): 340-357.

Bates, Douglas M., Mächler, Martin, Bolker, Ben & Walker, Steve. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1): 1-48.

Behaghel, Otto. 1909. Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern [Relationships between size and ordering of constituents]. *Indogermanische Forschungen* 25: 110-142.

Belsley, David A., Kuh, Edwin & Welsch, Roy E. 1980. *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: John Wiley.

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

Biber, Douglas. 2012. Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory* 8(1): 9-37. https://doi.org/10.1515/cllt-2012-0002

Biber, Douglas. 2019. Text-linguistic approaches to register variation. *Register Studies* 1(1): 42-75.

Biber, Douglas & Conrad, Susan. 2019. *Register, genre, and style*. 2nd ed. Cambridge: Cambridge University Press.

Biber, Douglas & Egbert, Jesse. 2018. *Register variation online*. Cambridge: Cambridge University Press.

Biber, Douglas, Egbert, Jesse, Gray, Bethany, Oppliger, Rahel & Szmrecsanyi, Benedikt. 2016. Variationist versus text-linguistic approaches to grammatical change in English: Nominal modifiers of head nouns. In *The Cambridge Handbook of English Historical Linguistics*, Merja Kytö & Päivi Pahta (eds), 351-375. Cambridge: Cambridge University Press.

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

Branigan, Holly P., Pickering, Martin J. & Tanaka, Mikihiro. 2008. Contributions of animacy to grammatical function assignment and word order during production. *Lingua* 118(2): 172-189. https://doi.org/10.1016/j.lingua.2007.02.003

Bresnan, Joan, Cueni, Anna, Nikitina, Tatiana & Baayen, R. Harald. 2007. Predicting the dative alternation. In *Cognitive Foundations of Interpretation*, Gerlof Boume, Irene Kraemer & Joost Zwarts (eds), 69-94. Amsterdam: Royal Netherlands Academy of Arts and Sciences.

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

Bresnan, Joan & Hay, Jennifer. 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, Herman, Ondřej, Jakubíček, Miloš, Krek, Simon & Novak, Blaž. 2017. *JSI newsfeed corpus*. Paper presented at the 9th International Corpus Linguistics Conference, University of Birmingham.

Bybee, Joan L. 2006. From usage to grammar: The mind's response to repetition. *Language* 82(4): 711-733.

Chafe, Wallace L. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In *Subject and topic*, Charles N. Li (ed), 25-56. New York: Academic Press

D'Arcy, Alexandra & Tagliamonte, Sali A. 2015. Not always variable: Probing the vernacular grammar. *Language Variation and Change* 27(3): 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/>.

Davies, Mark. 2018-. *The 14 billion word iWeb corpus*. <https://www.english-corpora.org/iWeb/>.

Davies, Mark & Fuchs, Robert. 2015. Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide*, 36(1): 1-28. https://doi.org/10.1075/eww.36.1.01dav

Ehmer, Oliver & Rosemeyer, Malte. 2018. When "questions" are not questions. Inferences and conventionalization in Spanish but-prefaced partial interrogatives. *Open Linguistics* 4: 70-100. https://doi.org/10.1515/opli-2018-0005

Ford, Marilyn & Bresnan, Joan. 2013. Using convergent evidence from psycholinguistics and usage. In *Research Methods in Language Variation and Change*, Manfred Krug & Julia Schlüter (eds), 295-312. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511792519.020

Garretson, Gregory, O'Connor, Catherine, Skarabela, Barbora & Hogan, Marjorie. 2004. Coding practices used in the project Optimality Typology of Determiner Phrases. Unpublished manuscript, Boston University.

Geleyn, Tim. 2017. Syntactic variation and diachrony: The case of the Dutch dative alternation. *Corpus Linguistics and Linguistic Theory* 13(1): 65-96. https://doi.org/10.1515/cllt-2015-0062

Gerwin, Johanna. 2014. *Ditransitives in British English dialects* [Topics in English Linguistics 50.3]. Berlin, Boston: De Gruyter Mouton.

Grafmiller, Jason. 2014. Variation in English genitives across modality and genres. *English Language and Linguistics* 18(3): 471-496. https://doi.org/10.1017/S1360674314000136

Grafmiller, Jason & Szmrecsanyi, Benedikt. 2018. Mapping out particle placement in Englishes around the world. A case study in comparative sociolinguistic analysis. *Language Variation and Change* 30(3): 385-412. https://doi.org/10.1017/S0954394518000170

Gries, Stefan Th. 2015. The most under-used statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora* 10(1): 95-125.

Grondelaers, Stefan, Speelman, Dirk & Geeraerts, Dirk. 2008. National variation in the use of *er* "there". Regional and diachronic constraints on cognitive explanations. In *Cognitive Sociolinguistics* [Cognitive Linguistics Research 39], Dirk Geeraerts, René Dirven, John R. Taylor, Ronald W. Langacker & Gitte Kristiansen (eds), 153-204. Berlin, New York: Mouton de Gruyter.

Gundel, Jeanette K., Hedberg, Nancy & Zacharski, Ron. 1993. Cognitive status and the form of referring expressions in discourse. *Language* 69(2): 274-307. https://www.jstor.org/stable /416535

Guy, Gregory R. 2005. Letters to *Language*. *Language* 81(3): 561-563.

Guy, Gregory R. 2015. *Coherence, constraints and quantities*. Paper presented at New Ways of Analyzing Variation (NWAV) 44, University of Toronto.

Hawkins, John A. 1994. *A performance theory of order and constituency*. Cambridge: Cambridge University Press.

Hayes, John R. & Flower, Linda S. 1980. Identifying the organisation of writing processes. In *Cognitive Processes in Writing*, Lee W. Gregg & Erwin Steinberg (eds), 3-30. Hillsdale: Lawrence Erlbaum Associates.

Heller, Benedikt, Szmrecsanyi, Benedikt & Grafmiller, Jason. 2017. Stability and fluidity in syntactic variation world-wide: The genitive alternation across varieties of English. *Journal of English Linguistics* 45(1), 3-27.

Jankowski, Bridget L. 2013. *A variationist approach to cross-register language variation and change*. PhD dissertation, University of Toronto.

Klavan, Jane & Divjak, Dagmar. 2016. The cognitive plausibility of statistical classification models: Comparing textual and behavioral evidence. *Folia Linguistica* 50(2), 355-384. https://doi.org/10.1515/flin-2016-0014

Koch, Peter & Oesterreicher, Wulf. 2012. Language of immediacy – Language of distance: Orality and literacy from the perspective of language theory and linguistic history. In *Communicative spaces*: *Variation, contact, and change*, Claudia Lange, Beatrix Weber & Göran Wolf (eds.), 441–473. Frankfurt: Lang.

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

Labov, William. 1984. Field methods of the project on linguistic change and variation. In *Language in Use: Readings in Sociolinguistics*, John Baugh & Joel Scherzer (eds), 28-53. Englewood Cliffs: Prentice Hall.

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

Levshina, Natalia. 2015. *How to do linguistics with R: Data exploration and statistical analysis*. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Love, Robbie, Dembry, Claire, Hardie, Andrew, Brezina, Vaclav & McEnery, Tony. 2017. Compiling and analysing 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 & Schuth, Anne. 2010. DutchParl: The parliamentary documents in Dutch. In *Proceedings of the Seventh International Conference on Linguistic Resources (LREC-2010)*, 3670-3677. European Language Resources Association.

Nakagawa, Shinichi & Schielzeth, Holger. 2013. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 4: 133-142.

Pijpops, Dirk & van de Velde, Freek. 2018. A multivariate analysis of the partitive genitive in Dutch. Bringing quantitative data into a theoretical discussion. *Corpus Linguistics and Linguistic Theory* 14 (1): 1–33. https://doi.org/10.1515/cllt-2013-0027.

R Core Team. 2019. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>

Rayner, Keith & Duffy, Susan A. 1986. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition* 14(3): 191-201.

Rickford, John R. 2014. Situation: Stylistic variation in sociolinguistic corpora and theory. *Language and Linguistics Compass* 8(11): 590-603. https://doi.org/10.1111/lnc3.12110

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

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

Röthlisberger, Melanie. 2018b. Guidelines for the dative alternation. Unpublished Manuscript.

Schönefeld, Doris. 2011. Introduction: On evidence and the convergence of evidence in linguistic research. In *Converging Evidence: Methodological and Theoretical Issues for Linguistic Research*, Doris Schönefeld (ed.), 1-31. Amsterdam: Benjamins. https://doi.org/10.1075/hcp.33.03sch

Shih, Stephanie & Grafmiller, Jason. 2011. *Weighing in on end weight*. Paper presented at the LSA 85th Annual Meeting, 6–9 January 2011, Pittsburgh, PA.

Szmrecsanyi, Benedikt. 2017. Variationist sociolinguistics and corpus-based variationist linguistics: Overlap and cross-pollination potential. *Canadian Journal of Linguistics/Revue Canadienne De Linguistique* 62(4): 685-701. https://doi.org/10.1017/cnj.2017.34

Szmrecsanyi, Benedikt. 2019. Register in variationist linguistics. *Register Studies* 1(1): 76-99. https://doi.org/10.1075/rs.18006.szm

Szmrecsanyi, Benedikt, Grafmiller, Jason, Bresnan, Joan, Rosenbach, Annette, Tagliamonte, Sali & Todd, Simon 2017. Spoken syntax in a comparative perspective: The dative and genitive alternation in varieties of English. *Glossa: A Journal of General Linguistics* 2(1): 1-27. http://doi.org/10.5334/gjgl.310

Tagliamonte, Sali A. 2013. Comparative sociolinguistics. In *Handbook of Language Variation and Change* (2nd edn), J. K. Chambers & Natalie Schilling (eds), 130–156. Chichester, UK: John Wiley & Sons, Ltd.

Tagliamonte, Sali A. 2016. So sick or so cool? The language of youth on the internet. *Language in Society* 45(1): 1-32.

Tagliamonte, Sali A., & D'Arcy, Alexandra. 2009. Peaks beyond phonology: Adolescence, incrementation, and language change. *Language* 85(1): 58-108.

Theijssen, Daphne, Bosch, Louis ten, Boves, Lou, Cranen, Bert & van Halteren, Hans. 2013. Choosing alternatives: Using Bayesian networks and memory-based learning to study the dative alternation. *Corpus Linguistics and Linguistic Theory* 9(2): 227-262. https://doi.org/10.1515/cllt-2013-0007

Wolk, Christoph, Bresnan, Joan, Rosenbach, Anette & Szmrecsanyi, Benedikt. 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

Wurm, Lee H. & Fisicaro, Sebastiano A. 2014. What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language* 72: 37-48. https://doi.org/10.1016/j.jml.2013.12.003

York, Richard. 2012. Residualization is not the answer: Rethinking how to address multicollinearity. *Social Science Research* 6(41): 1379-1386. https://doi.org/10.1016/j.ssresearch.2012.05.014

Zaenen, Annie, Carletta, Jean, Garretson, Gregory, Bresnan, Joan, Koontz-Garboden, Andrew, Nikitina, Tatiana, O'Connor, Catherine, & Wasow, Tom. 2004. Animacy encoding in English: Why and how. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, Barcelona, July 2004, Bonnie Webber & Donna Byron (eds), 118-125. East Stroudsburg, PA: Association for Computational Linguistics.

Zehentner, Eva. 2019. *Competition in language change: The rise of the English dative alternation* [Topics in English Linguistics 103]. Berlin, Boston: De Gruyter.

Zuur, Alain F., Ieno, Elena N., Walker, Neil, Saveliev, Anatoly A. & Smith, Graham M. 2009. *Mixed effects models and extensions in ecology with R*. New York, NY: Springer Science & Business Media.