A variationist perspective on the comparative complexity of four registers at the intersection of mode and formality

Abstract: In this paper, we operationalize register differences at the intersection of formality and mode, and distinguish four broad register categories: spoken informal (conversations), spoken formal (parliamentary debates), written informal (blogs), and written formal (newspaper articles). We are specifically interested in the comparative probabilistic/variationist complexity of these registers – when speakers have grammatical choices, are the probabilistic grammars regulating these choices more or less complex in particular registers than in others? Based on multivariate modeling of richly annotated datasets covering three grammatical alternations in two languages (English and Dutch), we assess the complexity of probabilistic grammars by drawing on three criteria: (a) the number of constraints on variant choice, (b) the number of interactions between constraints, and (c) the relative importance of lexical conditioning. Analysis shows that contrary to theorizing in variationist sociolinguistics, probabilistic complexity differences between registers are not quantitatively simple: formal registers are consistently the most complex ones, while spoken registers are the least complex ones. The most complex register under study is written-formal quality newspaper writing. We submit that the complexity differentials we uncover are a function of acquisitional difficulty, of on-line processing limitations, and of normative pressures.

Keywords: Dutch; English; probabilistic grammar; register; variation

*Corresponding author: Benedikt Szmrecsanyi, KU Leuven, Leuven, Belgium, E-mail: benszm@kuleuven.be. https://orcid.org/0000-0001-8844-6602
Alexandra Engel, KU Leuven, Leuven, Belgium. https://orcid.org/0000-0002-5890-3478

Open Access. © 2022 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.
1 Introduction

This paper marries the study of language complexity to register analysis via the variationist method. In what follows, we discuss each of these research orientations in turn. The study of language complexity has a long history in applied linguistics and Second Language Acquisition (SLA) research, where researchers typically consider complexity a gauge for learners’ proficiency in the target language, a descriptor for performance, and an index to benchmark development (see Ortega 2012 for discussion). But more recently, language complexity has also attracted attention in more theoretically oriented fields such as cross-linguistic typology, dialectology, and sociolinguistics (see Ehret et al. 2021 for a recent literature review). We now know that human languages, and varieties/dialects of the same language, can differ in their complexity – for example, the grammars of creole languages are demonstrably simpler than the grammars of “older” languages (McWhorter 2001). Complexity differentials across languages or dialects are theoretically interesting because they cannot have biological or communicative reasons: human beings – whatever their language background – are endowed with the exact same linguistic capacities, and especially in elementary speech situations, languages tend to have similarly complex communicative functions.

With all that being said, a crucial theme in the empirical literature on language complexity concerns the question of how to best measure complexity. Many analysts distinguish between measures of absolute complexity and measures of relative complexity (Miestamo 2009). Relative complexity measures boil down to difficulty, for example, for language users and language learners (Kusters 2003; Trudgill 2001). Absolute complexity measures (also known as “objective” measures) include quantitative complexity (e.g., the length of the minimal description of a linguistic system; Dahl 2004), redundancy-induced complexity (a.k.a. “baroque accretion”; see McWhorter 2001: 126), or irregularity-induced complexity (see e.g., Nichols 2013).

We have mentioned before the mantra that in elementary speech situations, languages tend to have similarly complex communicative functions. But of course not all speech situations are (equally) elementary, and so one would naturally expect that situational language varieties, a.k.a. registers, differ in complexity. This is why in the field of register studies, the issue of (comparative) complexity within and across registers (with a particular focus on specialized registers, such as academic prose) has received considerable attention as well. In register studies, the dominant approach is text-linguistic in nature, and as such uses methodologies such as Multidimensional (MD) analysis (see, e.g., Biber 1988; Biber and Conrad 2012). MD analysis is designed to investigate the functional relationship
between linguistic variation and the situational context: in a bottom-up fashion, registers are identified based on co-occurrence patterns of linguistic features in individual corpus texts. Building on the sizable text-linguistic literature utilizing MD analysis, the Register-Functional (RF) approach to grammatical complexity (see the papers in Biber et al. 2021a) draws inspiration from the fact that variation is inherent in human language, and that text-linguistic variation has a functional-contextual basis. But unlike classical complexity research (e.g., in SLA research and in applied linguistics), the RF approach eschews the customary notion that textual complexity is proportional to structural length and elaboration (see, in particular, Biber et al. 2011).

Similar to RF approaches, the study we report in this paper avoids simplistic complexity measures such as structural length and elaboration. But unlike RF approaches, we will not be engaging in text-linguistic analysis either. Rather, we seek new horizons by exploring the intersection between register studies, complexity analysis, and variationist linguistics. Variationist linguistics takes an interest in variation between “alternate ways of saying ‘the same’ thing” (Labov 1966: 188). The focus is therefore on intra-speaker variability (or “variability in the linguistic signal within a given language”, in the parlance of van Hout and Muysken 2016: 250). Variationist linguists carefully account for competing variants and use quantitative (typically multivariate/multifactorial) methodologies to model the probabilistic factors that constrain how language users choose between semantically and functionally equivalent variants (for seminal work, see Labov 1969; Gries 2003; Bresnan et al. 2007). Crucially, the basic unit of analysis in variationist linguistics is individual linguistic choices (and not texts, as in text-linguistic approaches). This is another way of saying that variationist linguists are primarily interested in the probabilistic conditioning of linguistic choices (and not so much in text frequencies, as in text linguistics). For a more detailed discussion of the differences between text linguistics and variationist linguistics, see Biber et al. (2016) and Szmrecsanyi (2019).

Probabilistic grammars (a.k.a. “variable grammars” in variationist sociolinguistics parlance), then, refer to the set of constraints and their probabilistic effects on how people choose between variants of a particular alternation/variable (Szmrecsanyi et al. 2019: 2). We exemplify with the well-known dative alternation in English. To encode dative relations, speakers and writers of English may use two semantically and functionally roughly equivalent structural patterns: the ditransitive dative variant, as in for example, someone gives you a DVD, and the prepositional dative variant, as in someone gives a DVD to you (see Section 2.1 for more discussion and exemplification). Bresnan et al. (2007) explore the conditioning factors that constrain language users’ dative choices in American English (telephone) conversations, as sampled in the Switchboard corpus of US American
English (Godfrey et al. 1992). Bresnan et al. find that in that particular register and regional variety, variation between the two dative options is constrained by about 10 probabilistic constraints, including, for example, pronominality of the recipient/theme, discourse accessibility (pragmatics), constituent length, and animacy of the recipient. If, for example, the recipient is inanimate (as in the browser sends a connection request to that site) instead of animate (as in someone gives you a DVD), 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. The combined effects of all relevant constraints, then, constitutes the probabilistic grammar of the dative alternation.

Previous work indicates that probabilistic grammars, for instance of the dative alternation in English, may be register-specific, in that constraints may have different effects in different registers (Engel et al. 2021; Szmrecsanyi and Engel 2021). In this paper, we investigate if register-specific probabilistic grammars also differ in terms of complexity. We are specifically going to be interested in the comparative probabilistic complexity of registers, and take a variationist approach to conceptualizing complexity. When speakers have grammatical choices, are the probabilistic grammars regulating these choices more or less complex in particular registers than in others?

This paper operationalizes register differences as follows. We investigate four registers defined at the intersection of formality and mode in line with Koch and Oesterreicher (2012). Formality is defined by situational characteristics such as the setting (private vs. public), topic/communicative purpose, and relationship between participants. In this view, formality constitutes a continuum, but note that the registers chosen here do not necessarily represent endpoints of this continuum. We specifically cover the following registers:

- **Spoken informal**: conversations between family members and friends
- **Spoken formal**: parliamentary debates
- **Written informal**: blogs/chats
- **Written formal**: quality newspaper articles

On the linguistic side, as case studies we will be studying three grammatical alternations: the dative alternation in English (someone gives you a DVD ∼ someone gives a DVD to you); the future temporal reference alternation in English (alien life forms we will come across ∼ alien life forms we are going to come across); and the dative alternation in Dutch (ik geef Sophieke ne friet ∼ ik geef ne friet aan Sophieke).

More information about these alternations will be provided in Section 2.

On the operational plane, we assess the complexity of probabilistic grammars by drawing on three criteria: (a) the number of constraints (e.g., recipient animacy
in the dative alternation) on variant choice, (b) the number of interactions between constraints (e.g., in the dative alternation, does recipient animacy have a stronger effect when the recipient is comparatively short?), and (c) the relative importance of lexical conditioning (e.g., in the dative alternation, do particular recipient lemmas strongly favor either dative variant?). The probabilistic complexity measures we utilize here pertain to absolute as well as relative complexity. For one thing, increases in the number of constraints, in the number of interactions between the constraints, and in the importance of lexical conditioning will increase the length of the description of probabilistic grammars, thus adding to absolute complexity. At the same time, it is plausible that absolutely more complex (that is, lengthier) grammars are, all other things being equal, more difficult to acquire and handle for language users, thus increasing relative complexity (see, e.g., Audring 2017: fn1).

The rationale behind our investigation is that the comparative probabilistic complexity of registers is largely unexplored, but has important implications for theorizing in variation studies and beyond. The customary view in variationist sociolinguistics is that stylistic variation – and, by extension, register variation – are “quantitatively simple” (Guy 2005: 562; see also e.g., Labov 2010: 265; Rickford 2014: 596). This is another way of saying that probabilistic grammars should be fairly identical across registers. Hence, major complexity differentials across stylistic and (again, by extension) situational varieties are not predicted. In this paper, we rely on multivariate/multifactorial analysis (specifically, mixed-effects logistic regression analysis and conditional inference tree analysis) to empirically debate this prediction, the research question being: Is it true that probabilistic complexity does not vary substantially as a function of register? Analysis shows that probabilistic complexity differences between registers are, in fact, not quantitatively simple.

This paper is structured as follows. Section 2 will discuss the alternations, methods, and data sources. Section 3 explains the analysis techniques and complexity criteria that this study relies on. Section 4 presents the results. In Section 5, we discuss our findings against the backdrop of what we know about register and complexity variation. Section 6 offers some concluding remarks.

2 Alternations and data

In this paper, we investigate three alternation patterns, two in English and one in Dutch. For English, we cover the dative alternation and the future temporal reference alternation. For Dutch, we will investigate the dative alternation, which is formally rather parallel to the dative alternation in English. We would like to
emphasize that in the context of the present paper, the selection of alternations subject to study is not a conceptual issue. As in Biber (1988)-inspired research, we use alternations as features to assess complexity differentials across registers. That being said, the rationale behind the selection of these particular alternations is the following. There is a huge variationist literature on both the dative alternation in English and the future temporal reference alternation in English, so that we can draw on a sizable body of literature. Note also that the two alternations cover different types of grammatical variation: The dative alternation is essentially a constituent order alternation, while the future temporal reference alternation is a substitution alternation in which different modal verb constructions alternate. Further, we study variation in English because register differences in English are comparatively well understood – consider that much previous MD and RF work is concerned with English. The case study on the Dutch dative alternation we include as a cross-linguistic plausibility check: Do complexity differences in Dutch, a language that is closely related to English, resemble those that we see in English? The dative alternation is a suitable target phenomenon in this context because of its formal and functional similarity in both English and Dutch.

In the remainder of this section, we sketch the alternations and datasets subject to analysis. Detailed dataset documentations are available as supplementary materials at https://osf.io/gdmtr/.

2.1 The dative alternation in English

In English, there are two ways of expressing dative relations: the ditransitive dative construction (recipient-theme order), and the prepositional dative construction (theme-recipient order), as in (1):

(1) If like someone gives [you]recipient [a DVD or a box set of something]theme then it’s like you’ve gotta watch it cos they’ve given [it]theme to [you]recipient and they expect that you give it back having fully watched it (Spoken BNC2014, SNXG)

This alternation is very well understood. In their seminal study, Bresnan et al. (2007) – discussed in the Introduction section – identify 10 probabilistic constraints on the dative alternation. Recent work has demonstrated that probabilistic constraints such as animacy and length have been subject to weakening in real time (Wolk et al. 2013), while the effects of pronominality and length differ across regional varieties (Röthlisberger et al. 2017). Importantly, there is also preliminary evidence that the effect of theme definiteness differs between spoken and written English (Theijssen et al. 2013), suggesting that probabilistic grammars may be register-specific.
2.1.1 Corpus materials

We tap into the following data sources:

- **Spoken informal**: conversations between family members and friends as provided in the Spoken BNC 2014 (~11.4 million words) (Love et al. 2017); private setting, private topics and various communicative purposes (Biber et al. 2021a), mostly close relationships between the participants.
- **Spoken formal**: parliamentary debates from the House of Commons (~59.4 million words) (Marx and Schuth 2010); public/institutionalized setting, public topics, persuasive and informative purposes (see Ilie 2015), (mostly) distant relationships between the participants (members of own party or opposition party, wider public).
- **Written informal**: British English blogs part of the GloWbE (~148 million words) (Davies and Fuchs 2015); public setting, private and public topics, various communicative purposes (Biber and Egbert 2018), mixed audience: writers and readers may or may not know each other closely.
- **Written formal**: online newspaper articles which have been scraped from the websites of The Independent (~113.5 million words) (Bušta et al. 2017); public setting and topics, informative purpose, distant relationships between writer(s) and reader(s).

We randomly sampled 650 variable tokens of the dative alternation with give per register, yielding a dataset of \( N_{\text{total}} = 2,600 \) observations. This dataset had a balanced distribution of dative variants per register: from each corpus, we included an equal number of ditransitive dative and prepositional dative tokens. Tokens were automatically extracted and sampled, manually checked for the variable context, and manually annotated for the constraints subject to analysis.

2.1.2 Defining the variable context

For reasons of space, we keep the description of the variable context short and to the point. Consistent with previous research (Bresnan et al. 2007; Röthlisberger et al. 2017; Theijssen et al. 2013), we excluded instances of non-canonical word order, particle verbs, passivized and relativized constituents, clausal or gerundial constituents, and fixed expressions. The full definition including exemplification can be found in the supplementary materials, available at https://osf.io/gdmtr/.

2.1.3 Annotation for probabilistic constraints

In what follows, we give a short overview of the constraints subject to analysis. Again, we refer the reader to the documentation for more detail (https://osf.io/gdmtr/).
Most of the following probabilistic constraints (except for verb sense, which is not constituent-specific) were annotated for both the recipient and the theme constituents separately.

- **Animacy**: animate (humans, animals) versus inanimate (incl. collective nouns), as in (2):

\[(2) \quad \text{I am not in a position to give [the honourable Gentleman]_{animate} [that information]_{inanimate}.} \quad \text{(House of Commons, 2009-11-12)}\]

- **Pronominality**: pronominal versus non-pronominal, as in (3):

\[(3) \quad \text{The concept of Freecycling allows a person to give [his unwanted stuff]_{non-pronominal} to [someone who deserves it]_{pronominal}, cutting back on wastage} \quad \text{(GloWbE-GB, blogs)}\]

- **Definiteness**: definite versus indefinite, as in (4):

\[(4) \quad \text{They certainly don’t give [these people]_{definite} [any training]_{indefinite} (Spoken BNC2014, SYNT)}\]

- **Givenness**: given versus new, as in (5):

\[(5) \quad \text{Yeah, I mean it gives [you]_{given} [a proper basis for life]_{new}, you know} \quad \text{(Spoken BNC2014, SEPP)}\]

- **Complexity**: simple versus complex (post-modification of the head), as in (6):

\[(6) \quad \text{She spoke about the importance of testimony from victims, and of giving [a voice]_{simple} to [those who have none]_{complex}.} \quad \text{(The Independent, 2017-09-17)}\]

- **Constituent Length**: measured in number of characters; a weight ratio was calculated by dividing the length of the recipient by the length of the theme

- **Frequency**: register-specific frequency of head lemma per million words

- **Thematicity**: frequency of head lemma divided by number of words in a text

- **Verb Sense**: transfer (7a) versus abstract (7b) versus communication (7c)

\[(7) \quad \text{a. transfer: A foolish mistake and on returning to my car the police had arrived and given me a ticket.} \quad \text{(House of Commons, 2010-01-18)}\]

\[\text{b. abstract: It always seems like your schedule is constantly filling up, without an individual even giving it an additional thought.} \quad \text{(GloWbE-GB, blogs)}\]

\[\text{c. communication: Can you give me a definition?} \quad \text{(Spoken BNC2014, SH2P)}\]
2.2 The future temporal reference alternation in English

Future temporal reference (henceforth: FTR) in English is overtly expressed by means of either the modal verb will, or the semi-modal construction be going to. Consider (8):

(8) In practice, however, experts think the most likely alien life forms we will come across are going to be some kind of alien microbes. (The Independent, 2018-07-02)

The variationist literature suggests that the choice between will and be going to is conditioned by factors such as clause type and sentence type: be going to is favored in subordinate clauses – especially protases – and interrogative sentences, while will is favored in main clauses and declarative sentences (e.g., Denis and Tagliamonte 2018; Tagliamonte et al. 2014; Torres Cacoullos and Walker 2009). Other predictors, such as proximity of future time reference or polarity are more variable and may be subject to grammaticalization processes (Tagliamonte et al. 2014). As to register differences, we know that be going to is more common in informal language use than in formal language use (see, e.g., Mair 1997).

2.2.1 Corpus materials

Tokens were retrieved from the same corpus materials as in our study on the English dative alternation (see Section 2.1.1). Again, we created a random balanced sample of $N_{\text{total}} = 2,600$ observations of will and be going to.

2.2.2 Defining the variable context

Following Denis and Tagliamonte (2018), we excluded be going to in past tense contexts, and will in tag-questions (for more details, see the supplementary materials at https://osf.io/gdmtr/).

2.2.3 Annotation for probabilistic constraints

A detailed description is available in the supplementary materials at https://osf.io/gdmtr/.

- **Sentence Type**: Declarative sentences (9a) versus interrogative sentences (9b)

(9) a. declarative: Facebook said it will double its safety and security staff to 20,000, including contract workers, by the end of 2018. (The Independent, 2017-11-01)

b. interrogative: Am I going to celebrate today? (GloWbE-GB, blogs)
CLAUSE TYPE: Main clauses (10a) versus subordinate clauses (10b)

(10) a. main clause: The Crew Dragon capsule will remain attached to the ISS until 8 March, when it will detach and land back on Earth. (The Independent, 2019-02-28)

b. subclause: Unless things change that’s what I’m going to have to do (BNC2014, S59W)

We also distinguish between apodosis (11a) versus protasis (11b) in conditional clauses versus non-conditional clauses (11c)

(11) a. apodosis: If you put that in the washing machine it’ll be fine (BNC2014, SUVQ)

b. protasis: It should be proof that if mitochondrial treatment is going to take place then doctors should begin as soon as they can, they said. (The Independent, 2017-07-24)

c. non-conditional: And we will not cut spending as has been made clear by the Government. (House of Commons, 2008-10-21)

POLARITY: affirmative (12a) versus negative (12b)

(12) a. affirmative: I’ll do my best to ensure that I know what I’m talking about. (House of Commons, 2013-01-09)

b. negative: Can’t see now, so I’m never gonna see again (BNC2014, STSS)

GRAMMATICAL PERSON OF THE SUBJECT: First-person (13a) versus second-person (13b) versus third-person subjects (13c)

(13) a. first person: It looks excellent and well worth reading, I am sure I could learn from it so I am going to buy it! (GloWbE-GB, blogs)

b. second person: You’ll have to de-stress (BNC2014, S6YA)

c. third person: Ofcom is going to take up the cudgels, and I hope it will do that sooner rather than later. (House of Commons, 2011-05-19)

ANIMACY OF THE SUBJECT: animate (14a) versus inanimate (14b)

(14) a. animate: But he’ll see you’ve unfriended him (BNC2014, SCMZ)

b. inanimate: The 15-year contracts set an agreed minimum price that electricity generators will receive for supplying electricity. (The Independent, 2018-05-16)

PROXIMITY OF FUTURE TEMPORAL REFERENCE: proximate (15a) versus distal (15b) versus no reference contexts (15c)
(15) a. proximate: I think in terms of policy statement, that is as close as we are going to get tonight. (House of Commons, 2011-04-04)
   b. distal: We’re also going to talk this week at FP Day in Cambridge. (GloWbE-GB, blogs)
   c. no reference: The government has faced sustained pressure over fears about the impact the policy will have on millions of low-paid people. (The Independent, 2018-11-04)

- Presence of temporal adverbial: present (16a) versus absent (16b) in the same clause

(16) a. present: I’m gonna go to work tomorrow and ring the tax office about my tax code (BNC2014, SXCW)
   b. absent: How robust is it going to be in that case? (The Independent, 2017-12-07)

- Lexical verb type: dynamic/non-motion (17a) versus motion (17b) versus stative verbs (17c)

(17) a. dynamic: And I will say more of this as I go through my speech. (House of Commons, 2008-07-03)
   b. motion: Yeah, so I anticipate really in the next ten years a lot of people will come back home (BNC2014, S9YG)
   c. stative: First of all, ask yourself what the average customer is going to want to see from a business like your own. (GloWbE-GB, blogs)

2.3 The dative alternation in Dutch

Similar to English, there are two competing dative constructions in Dutch: the ditransitive dative construction, and the prepositional dative construction with aan. Consider (18):

(18) Ey as[sic!] atari [ ons]_recipient [ gene friet]_theme geeft dan geef ik Ey if Atari us no fries-SG gives then give I [ ne friet]_theme aan [sphieke[sic!]]_recipient se a fries-SG to Sophieke see
   ‘Hey, if Atari doesn’t give us fries, then I’ll give fries to Sophieke, you know’
   (SoNaR, chats, WRUEA_UN_190)
Most previous research on the Dutch dative alternation is not particularly data-driven. One notable exception is Geleyn (2017), who on the basis of a corpus consisting of novels demonstrates that the Dutch dative alternation is governed by largely the same predictors as the English dative alternation.

2.3.1 Corpus materials

The aim was to select corpora that are maximally comparable to those investigated for English. In this endeavor, we restricted attention to Belgian Dutch (as opposed to Netherlandic Dutch).

- **Spoken informal:** informal conversations between family members and friends (face-to-face and via telephone), and informal interviews with teachers of Dutch as provided in the Spoken Dutch Corpus (≈2 million words) (Oostdijk 2000); private setting, private topics and various communicative purposes (Biber et al. 2021a), mostly close relationships between the participants

- **Spoken formal:** parliamentary debates from the Flemish parliament (≈9.2 million words) (Marx and Schuth 2010); public/institutionalized setting, public topics, persuasive and informative purposes (Ilie 2015), (mostly) distant relationships between the participants (members of own party or opposition party, wider public)

- **Written informal:** chats from the Dutch SoNaR corpus (≈10 million words) (Oostdijk 2000); public setting, private and public topics, various communicative purposes, mixed audience: writers and readers may or may not know each other closely

- **Written formal:** online newspaper articles which have been scraped from the websites of De Morgen (≈22 million words) (Bušta et al. 2017); public setting and topics, informative purpose, distant relationships between writer(s) and reader(s)

We sampled tokens of both variants from these corpora. Due to sparse data in the informal corpora, the resulting dataset of \( N_{\text{total}} = 2,110 \) tokens is not perfectly balanced across registers: there are only 220 tokens (189 ditransitive dative, 31 prepositional dative) from the spoken informal register, and 590 tokens (325 ditransitive dative, 265 prepositional dative) from the written informal register, while 650 tokens are included from both formal registers. Keeping in mind the customary rule of thumb that the maximal number of parameters in a regression model should be calculated as the frequency of the less frequent outcome divided by 10 (Hosmer and Lemeshow 2000: 346–347), this means that the regression model covering the spoken informal register will be unreliable. We will be flagging this issue in Sections 4 and 5.
2.3.2 Defining the variable context

The variable context was defined in the same way as for the English dative alternation. The full definition, including exemplification, is available in the supplementary materials at https://osf.io/gdmtr/.

2.3.3 Annotation for probabilistic constraints

The predictor set matches that of the English dative alternation. For more details, see the supplementary materials at https://osf.io/gdmtr/.

- **ANIMACY**: animate (humans, animals) versus inanimate (including collective nouns), as in (19):

  (19) *Het was gewoon een manier om [mijn medemens]animate [een
godegevoel][sic!]]inanimate te geven
good feeling to give.
‘It was just a manner to make my fellow man feel good.’
(SoNaR, WRUEA_UN_150)

- **PRONOMINALITY**: pronominal versus non-pronominal, as in (20):

  (20) *Mijnheer Vandaele, zal ik [u]pronominal anders [mijn
nummer]non-pronominal geven?
number give?
‘Mr Vandaele, shall I give you my number instead?’
(Flemish Parliament, 2012-11-28)

- **DEFINITENESS**: definite versus indefinite, as in (21):

  (21) *Wie daar subnationaliteit in ziet, geeft [een andere onbedoelde
lezing]indefinite aan [de tekst]definite-
reading to the text.
‘Anyone who sees sub-nationality in it gives another unintended reading to the text.’
(Flemish Parliament, 2009-07-15)
- **Givenness**: given versus new, as in (22):

(22) *Als ik een klas van twintig leerlingen heb dan probeer ik het eerste trimester uhm [de leerlingen elk] given [een taak] new te geven.*  
‘If I have a class of 20 students, I try to give each student a task in the first trimester.’  
(Spoken Dutch Corpus, fv400152)

- **Complexity**: simple versus complex (post-modification of the head), as in (23):

‘Two leading child psychiatrists follow what happens and give explanations or good advice to parents.’  
(De Morgen, 2015-09-02)

- **Constituent length**: measured in number of characters; a weight ratio was calculated by dividing the length of the recipient by the length of the theme

- **Frequency**: register-specific frequency of head lemma per million words

- **Thematicity**: frequency of head lemma divided by number of words in a text

- **Verb sense**: transfer (24a) versus abstract (24b) versus communication (24c)

(24) a. transfer: *Ze geeft hem geld zodat hij water en extra eten kan kopen.*  
‘She gives him money so that he can buy water and extra food.’  
(De Morgen, 2016-08-20)

b. abstract: *Vooral ’t feit dat ze op kot zit geeft haar enorme vrijheid.*  
‘Especially the fact that she is living in digs gives her enormous freedom’  
(Spoken Dutch Corpus, fv701256)

c. communication: *Je moet dan goede informatie aan hem geven.*  
‘Then you have to give good information to him’  
(SoNaR, WRUEA_UN_109)
3 Analysis methods and complexity criteria

We will use two analysis methods in this paper. The first one, binary logistic regression analysis, is the workhorse analysis technique in variationist linguistics, and quantifies the simultaneous effect of multiple individual explanatory factors (or: constraints/predictors/independent variables) on a binary dependent variable, such as dative or future marker outcomes. We utilize a modern refinement of logistic regression analysis known as mixed-effects logistic regression (Pinheiro and Bates 2000). In addition to so-called fixed effects – which are classically estimated predictors suited for assessing predictor significance and effect directions – mixed-effects modeling also cover so-called random effects designed to capture variation dependent on open-ended, potentially hierarchical and unbalanced groups (see Wolk et al. 2013: 399–400 for more discussion). Specifically, we will include in our analysis so-called by-item random effects to measure lexical effects: what is the extent to which, for example, particular theme lemmas favor particular dative constructions? What is the extent to which particular lexical verbs favor particular future markers? This is information that mixed-effects regression models are ideally suited to provide.

In addition, we will be conducting a series of conditional inference tree analyses. Szmrecsanyi et al. (2016: 113–114) succinctly summarize the basic idea behind conditional inference trees as follows:

conditional inference trees […] predict outcomes by recursively partitioning the data into smaller and smaller subsets according to those predictors that co-vary most strongly with the outcome. Informally, binary splits in the data are made by trying to maximize the homogeneity or “purity” of the data partitions with respect to the values of the outcome (e.g., all s-genitives vs. all of-genitives). At each step, the dataset is recursively inspected to determine the variable (and its values) that yields the purest split in the data. This splitting process is repeated until no further split that significantly reduces the impurity of the data partitions can be found. The result is visualized as a flowchart-like decision tree.

Crucially, each split can be interpreted as an interaction between predictors: the more one moves downward a tree, the more the effect of particular predictors depends on predictors further up in the tree. For an accessible introduction to conditional inference tree analysis from a linguistic angle, see Tagliamonte and Baayen (2012). Conditional inference trees are therefore extremely convenient and elegant tools to detect interaction effects in a bottom-up fashion, without the model fitting/model selection/model pruning complexities that regression models would incur.

Based on output from the aforementioned analysis methods, we subsequently use three criteria (or: metrics) to assess the complexity of probabilistic grammars:
1. **Sheer number of constraints on variation**: This is the one probabilistic complexity criterion that has actually been proposed in the literature (Shin 2014): the number of constraints on variation is proportional to probabilistic complexity. Hence, a variety or register where, say, the dative alternation is constrained by 10 factors is probabilistically more complex (with regard to that alternation!) than a variety/register where we find only five significant constraints. The rationale is that (a) as per absolute-quantitative complexity definitions (Dahl 2004), more constraints require more description; and (b) probabilistic grammars involving more constraints are arguably harder to acquire and so more complex, as per the relative complexity criterion (Kusters 2003). We operationalize this criterion by counting the number of significant constraints in register-specific regression models.¹

2. **Number of interactions between constraints**: Even a small set of constraints can engender complexity if the constraints interact robustly. For example, the dative alternation may be constrained by no more than, say, three constraints in some hypothetical register, but three constraints may still give rise to up to three 2-way interactions and one three-way interaction, which yields a fairly complex constraint system after all. The rationale is that as per the definition of absolute-quantitative complexity, interactions require more description, and hence induce complexity. We operationalize this criterion by counting the number of nodes in conditional inference trees.

3. **Relative importance of lexical conditioning**: Some constraints on variation are lexical in nature. For example, it is well known that particular dative verbs may be biased towards particular dative variants (Gries and Stefanowitsch 2004). Our assumption that lexical constraints induce probabilistic complexity is primarily motivated by relative complexity considerations: consider that while end weight effects, which are likely to be rooted in processing, seem to be acquired relatively early (Tanaka 1987), verb bias in the dative alternation appears to be observable only in advanced learners (Wolk et al. 2011), presumably because the acquisition of verb bias requires sufficient exposure to the

¹ We acknowledge that the sheer number of constraints could probably also be extracted from conditional inference tree models (although isolating main effects from interaction effects would pose challenges). Because we need regression models anyway to assess criterion #3 (lexical conditioning), we choose to infer the sheer number of constraints from regression models, as in Shin (2014). We would further like to acknowledge that the identification of significant constraints relies on an interpretation of p-values at the customary alpha level of 0.05. Constraints that score p-values greater than 0.05 will not count as significant constraints. Needless to say, absence of statistical significance does not equate with evidence of absence of substantial significance. The full label for criterion #1 is therefore “sheer number of constraints that pass the customary significance threshold of alpha = 0.05”.
specific lexical items. Further, as regards absolute complexity it is clear that lexical conditioning will increase the length of the shortest possible description of probabilistic grammars. We operationalize this criterion by including in mixed-effects regression modeling by-item lexical random effects, and subsequently gauging the explanatory power of these random effects.

It should be noted here that from a purely statistical point of view, criterion 1 (sheer number of constraints) and criterion 2 (number of interactions) both address quantitative model complexity: in conjunction, these criteria gauge the number of parameters in multivariate models. However, on the interpretational plane, main effects and interactions strike as conceptually distinct in nature, which is why we would like to keep them apart analytically.

In all, then, our results will be based on 24 multivariate models (12 regression models and 12 conditional inference trees – four each per alternation). This comparatively large number of models may strike some as excessive at first glance. Consider, however, that we investigate three alternations in four registers with regard to three complexity criteria. Consider also that text-linguistic studies on register in the spirit of Biber (1988) are based on a frequency analysis of dozens of features in potentially hundreds of texts, which can easily sum up to thousands of frequency measurements – and few would argue that this rich empirical basis is excessive. Because in this paper we engage in variationist/probabilistic linguistics, we need probabilistic measurements, and these (unlike frequency measurements) require prior multivariate modeling. This is why the number of models we calculate is proportionate, given our research objectives.

4 Results

This section reports the multivariate models upon which our complexity analysis is based. Specifically, for each of the three alternations, we report register-specific regression models (relevant to complexity criteria #1 and #3; see previous Section) and register-specific conditional inference trees (complexity criterion #2). Because of the comparatively large number of models we report, we will keep the summaries below short and to the point.
4.1 The dative alternation in English

4.1.1 Regression analysis

Table 1 indicates that the regression model for the written formal register includes seven significant constraints, while the model for the written informal register includes six; regression analysis further shows that the model for the spoken informal register includes only five significant constraints, and that the model for the spoken formal register is the least complex one, with only four significant constraints. Interestingly, the four registers differ not only with regard to the number of constraints involved, but also in terms of the set of constraints. Apart from weight ratio, none of the other constraints are consistently included in all four probabilistic grammars.

As shown in Table 2, fixed (i.e. non-lexical) effects alone explain more variance in the two informal registers than in the two formal registers (as evidenced by the $R^2$-values). The $\sigma^2_{\text{ThemeLemma}}$ values indicate that the random effect for theme lemma covers most variance in the spoken informal register, and least variance in the spoken formal register.

**Table 1:** Overview of significant and non-significant constraints in the English dative alternation based on logistic mixed-effects regression analysis of four registers. Significance codes: ‘***’ – $p < 0.001$; ‘**’ – $p < 0.01$; ‘*’ – $p < 0.05$; ‘ns’ – $p \geq 0.05$. C-index values range between 0.5 and 1 and indicate the goodness-of-fit of the regression models: the closer to 1, the better the fit.

<table>
<thead>
<tr>
<th>Constraint</th>
<th>Spoken informal ($C = 0.99$)</th>
<th>Spoken formal ($C = 0.96$)</th>
<th>Written informal ($C = 0.97$)</th>
<th>Written formal ($C = 0.97$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recipient Animacy</td>
<td>ns</td>
<td>***</td>
<td>ns</td>
<td>**</td>
</tr>
<tr>
<td>Theme Animacy</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>Recipient Pronominality</td>
<td>***</td>
<td>**</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>Theme Pronominality</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>Recipient Definiteness</td>
<td>ns</td>
<td>ns</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>Theme Definiteness</td>
<td>ns</td>
<td>ns</td>
<td>**</td>
<td>ns</td>
</tr>
<tr>
<td>Recipient Givenness</td>
<td>*</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>Theme Givenness</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>Recipient Complexity</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>Theme Complexity</td>
<td>ns</td>
<td>ns</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>Weight Ratio (Length)</td>
<td>***</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>Recipient Frequency</td>
<td>ns</td>
<td>ns</td>
<td>***</td>
<td>**</td>
</tr>
<tr>
<td>Theme Frequency</td>
<td>**</td>
<td>ns</td>
<td>**</td>
<td>ns</td>
</tr>
<tr>
<td>Recipient Thematicity</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>Theme Thematicity</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>Verb Sense</td>
<td>ns</td>
<td>ns</td>
<td>**</td>
<td>***</td>
</tr>
</tbody>
</table>
4.1.2 Conditional inference tree analysis

Conditional inference tree analysis also reveals complexity differentials. By counting the sheer number of constraints, informal registers (Figures 1 and 2) seem to be less complex than formal registers (Figures 3 and 4).

Table 2: Relative importance of lexical conditioning in the English dative alternation. Lexical random effect (intercept adjustment) subject to analysis: **theme lemma** (**recipient lemma** did not turn out to have substantial explanatory power). $R^2_{marginal}$: % variance explained by fixed effects; $R^2_{conditional}$: % variance explained by both random and fixed effects; $\sigma^2$: mean random effect variance of the model; SD: standard deviation; $N$: number of theme lemmas distinguished in the model (infrequent lemmas were binned into an ‘other’ category).

<table>
<thead>
<tr>
<th></th>
<th>Spoken informal</th>
<th>Spoken formal</th>
<th>Written informal</th>
<th>Written formal</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2_{marginal}$</td>
<td>0.81</td>
<td>0.74</td>
<td>0.83</td>
<td>0.67</td>
</tr>
<tr>
<td>$R^2_{conditional}$</td>
<td>0.92</td>
<td>0.80</td>
<td>0.89</td>
<td>0.83</td>
</tr>
<tr>
<td>$\sigma^2_{theme lemma}$</td>
<td>4.11</td>
<td>1.15</td>
<td>1.84</td>
<td>3.11</td>
</tr>
<tr>
<td>SD</td>
<td>2.03</td>
<td>1.07</td>
<td>1.36</td>
<td>1.76</td>
</tr>
<tr>
<td>$N_{theme lemma}$</td>
<td>33</td>
<td>26</td>
<td>35</td>
<td>56</td>
</tr>
</tbody>
</table>

**Figure 1:** Conditional inference tree of the English dative alternation in the spoken informal register ($C = 0.97$).
Figure 2: Conditional inference tree of the English dative alternation in the spoken formal register ($C = 0.93$).

Figure 3: Conditional inference tree of the English dative alternation in the written informal register ($C = 0.95$).
4.2 The FTR alternation in English

4.2.1 Regression analysis

Results from regression models are summarized in Table 3. The model with the largest number of significant constraints – six to be exact – is the model for the spoken formal register. The model for the written formal register includes five significant constraints, while the models covering the informal registers include four significant constraints. As for the specific constraints involved, we observe some overlap between the spoken formal, written informal, and written formal registers.

Table 3: Overview of significant and non-significant constraints in the English FTR alternation based on logistic mixed-effects regression analysis of four registers. Significance codes: ‘****’ – \( p < 0.001 \); ‘***’ – \( p < 0.01 \); ‘**’ – \( p < 0.05 \); ‘*’ – \( p \geq 0.05 \). C-index values range between 0.5 and 1 and indicate the goodness-of-fit of the regression models: the closer to 1, the better the fit.

<table>
<thead>
<tr>
<th></th>
<th>Spoken informal (( C = 0.64 ))</th>
<th>Spoken formal (( C = 0.75 ))</th>
<th>Written informal (( C = 0.73 ))</th>
<th>Written formal (( C = 0.75 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>SENTENCE TYPE</td>
<td>***</td>
<td>ns</td>
<td>**</td>
<td>*</td>
</tr>
<tr>
<td>CLAUSE TYPE</td>
<td>***</td>
<td>ns</td>
<td>***</td>
<td>**</td>
</tr>
<tr>
<td>CONDITIONAL CLAUSE</td>
<td>*</td>
<td>**</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>POLARITY</td>
<td>ns</td>
<td>**</td>
<td>ns</td>
<td>***</td>
</tr>
<tr>
<td>GRAMMATICAL PERSON</td>
<td>ns</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>ANIMACY OF SUBJECT</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>PROXIMITY OF FUTURE</td>
<td>ns</td>
<td>**</td>
<td>*</td>
<td>***</td>
</tr>
<tr>
<td>TEMPORAL REFERENCE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRESENCE OF TEMPORAL</td>
<td>ns</td>
<td>*</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>ADVERBIAL</td>
<td></td>
<td>ns</td>
<td></td>
<td>ns</td>
</tr>
<tr>
<td>LEXICAL VERB TYPE</td>
<td>*</td>
<td>*</td>
<td>ns</td>
<td>ns</td>
</tr>
</tbody>
</table>

Figure 4: Conditional inference tree of the English dative alternation in the written formal register (\( C = 0.92 \)).
Table 4: Relative importance of lexical conditioning in the English FTR alternation. Lexical random effect (intercept adjustment) subject to analysis: LEXICAL VERB LEMMA. $R^2_{\text{marginal}}$: % variance explained by fixed effects; $R^2_{\text{conditional}}$: % variance explained by both random and fixed effects; $\sigma^2$: mean random effect variance of the model; SD: standard deviation; $N$: number of theme lemmas distinguished in the model (infrequent lemmas were binned into an ‘other’ category).

<table>
<thead>
<tr>
<th></th>
<th>Spoken informal</th>
<th>Spoken formal</th>
<th>Written informal</th>
<th>Written formal</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2_{\text{marginal}}$</td>
<td>0.10</td>
<td>0.21</td>
<td>0.17</td>
<td>0.20</td>
</tr>
<tr>
<td>$R^2_{\text{conditional}}$</td>
<td>0.10</td>
<td>0.29</td>
<td>0.21</td>
<td>0.24</td>
</tr>
<tr>
<td>$\sigma^2_{\text{lexical verb}}$</td>
<td>0 (ns)</td>
<td>0.39</td>
<td>0.17</td>
<td>0.18</td>
</tr>
<tr>
<td>SD</td>
<td>0 (ns)</td>
<td>0.63</td>
<td>0.41</td>
<td>0.43</td>
</tr>
<tr>
<td>$N$</td>
<td>36</td>
<td>37</td>
<td>49</td>
<td>33</td>
</tr>
</tbody>
</table>

Lexical conditioning seems to be most powerful in the spoken formal register, and least important in the spoken informal register, where the lexical random effect is actually not significantly improving model fit, as shown in Table 4. The power of lexical conditioning is about equal in both written registers.

4.2.2 Conditional inference tree analysis

Conditional inference tree analysis suggests that there are only two interactions (nodes) in the spoken informal register (Figure 5). In the two formal registers

![Conditional inference tree diagram](image)

**Figure 5:** Conditional inference tree of the English FTR alternation in the spoken informal register ($C = 0.60$).
Figure 6: Conditional inference tree of the English FTR alternation in the spoken formal register ($C = 0.67$).

Figure 7: Conditional inference tree of the English FTR alternation in the written informal register ($C = 0.69$).

Figure 8: Conditional inference tree of the English FTR alternation in the written formal register ($C = 0.68$).
(Figures 6 and 8), conditional inference tree analysis uncover four interactions each, while the conditional inference tree for the written informal register has five interactions (Figure 7).

4.3 The dative alternation in Dutch

4.3.1 Regression analysis

Mixed effects logistic regression analysis (Table 5) reveals that the models covering the informal registers both include five significant predictors, while the model for the spoken formal register includes six significant predictors. The model for the written formal register has seven significant predictors. We further note that WEIGHT RATIO and THEME DEFINITENESS are significant in all registers. However, THEME ANIMACY, THEME COMPLEXITY, THEME FREQUENCY, THEME THEMATICITY, and RECIPIENT GIVENNESS are not significant in any register.

Both RECIPIENT LEMMA and THEME LEMMA lexically condition the Dutch dative alternation, although THEME LEMMA accounts for more variance than RECIPIENT LEMMA. Across

<table>
<thead>
<tr>
<th></th>
<th>Spoken informal (C = 0.95) (unreliable)</th>
<th>Spoken formal (C = 0.92)</th>
<th>Written informal (C = 0.94)</th>
<th>Written formal (C = 0.92)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RECIPIENT ANIMACY</td>
<td>ns</td>
<td>*</td>
<td>ns</td>
<td>***</td>
</tr>
<tr>
<td>THEME ANIMACY</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>RECIPIENT PRONOMINALITY</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>THEME PRONOMINALITY</td>
<td>*</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>RECIPIENT DEFINITENESS</td>
<td>*</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>THEME DEFINITENESS</td>
<td>***</td>
<td>*</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>RECIPIENT GIVENNESS</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>THEME GIVENNESS</td>
<td>ns</td>
<td>**</td>
<td>ns</td>
<td>**</td>
</tr>
<tr>
<td>RECIPIENT COMPLEXITY</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>***</td>
</tr>
<tr>
<td>THEME COMPLEXITY</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>WEIGHT RATIO (LENGTH)</td>
<td>*</td>
<td>***</td>
<td>**</td>
<td>***</td>
</tr>
<tr>
<td>RECIPIENT FREQUENCY</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>THEME FREQUENCY</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>RECIPIENT THEMATICITY</td>
<td>*</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>THEME THEMATICITY</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>VERB SENSE</td>
<td>ns</td>
<td>***</td>
<td>*</td>
<td>***</td>
</tr>
</tbody>
</table>

Table 5: Overview of significant and non-significant constraints in the Dutch dative alternation based on logistic mixed-effects regression analysis of four registers. Significance codes: ‘****’ – \( p < 0.001 \); ‘***’ – \( p < 0.01 \); ‘**’ – \( p < 0.05 \); ‘*’ – \( p \geq 0.05 \). C-index values range between 0.5 and 1 and indicate the goodness-of-fit of the regression models: the closer to 1, the better the fit. Note that because of data sparsity, the model covering the spoken informal register is unreliable.
registers, the random effect for \textsc{recipient lemma} explains most variance in the spoken formal register, and least variance in the written formal register (see Table 6). The random effect for \textsc{theme lemma} explains most variance in the written informal register, and least variance in the spoken formal register. Lexical conditioning via \textsc{theme lemma} is generally more important in the written registers than in the spoken registers.

### 4.3.2 Conditional inference tree analysis

The conditional inference tree for the spoken informal register includes three interactions (Figure 9), the tree for the spoken formal register includes six interactions (Figure 10), and the trees for the written informal (Figure 11) and the written formal registers (Figure 12) include five and seven interactions, respectively.

### 5 Discussion

Recall that in Section 3, we defined three criteria to assess the complexity of register-specific probabilistic grammars:

1. Sheer number of constraints on variation
2. Number of interactions between constraints
3. Relative importance of lexical conditioning

In what follows, we evaluate the regression models and conditional inference trees reported in the preceding section with regard to these criteria.
We begin with the dative alternation in English. The relevant information is summarized in Table 7. The overall complexity ranking, based on all criteria, is displayed in (25): 2

Figure 9: Conditional inference tree of the Dutch dative alternation in the spoken informal register (C = 0.83).

Figure 10: Conditional inference tree of the Dutch dative alternation in the spoken formal register (C = 0.81).

We begin with the dative alternation in English. The relevant information is summarized in Table 7. The overall complexity ranking, based on all criteria, is displayed in (25): 2

2 In determining the overall complexity ranking, all criteria were weighted equally. Registers were assumed to have the same rank when the rank aggregation differentials were smaller than 0.5.
The fairly uncontested “winner” in terms of overall complexity is the written formal register. At the other end of the continuum, we find the spoken formal register, which is overall the least complex register (though we note that the spoken formal register does exhibit a fairly high number of interactions). Spoken informal and written informal take the middle road and are fairly similar in terms of overall complexity.

Information about the FTR alternation in English is provided in Table 8. The overall complexity ranking is shown in (26):

\[(25)\quad \text{Written formal} > \text{spoken informal} \approx \text{written informal} > \text{spoken formal}\]
Spoken formal and written formal are overall about equally complex, with spoken formal having a slight edge. The clearly least complex register is spoken informal. Written informal exhibits medium complexity.

Lastly, Table 9 assesses the dative alternation in Dutch. The overall complexity ranking is shown in (27):

\[(27) \text{Written formal} \approx \text{spoken formal} \approx \text{written informal} > (\text{spoken informal})\]

Written formal, spoken formal, and written informal all have about the same level of overall complexity, though written formal is slightly ahead of the other two. The spoken informal register is substantially less complex than the other registers. That said, we bracket the spoken informal register because of unreliability due to data sparsity.
In the big picture, then, we conclude that there is quite a bit of variability regarding the overall complexity rankings across case studies, in that the complexity hierarchies are far from identical. (In addition, of course, there is also a fair amount of inter-criterion variability in each case study.) With that being said, however, the following three generalizations emerge from the hierarchies in (25), (26), and (27):

1. Written formal is the most complex register in two of the three case studies covered in this paper (the dative alternation in English as well as in Dutch). In the English FTR case study, written formal and spoken formal actually share the top complexity rank.
2. When we restrict attention to the endpoints of the hierarchies, consistently across the three case studies formal registers are the most complex ones.
3. Also consistently across the three case studies, spoken registers are the least complex ones.

What does it all mean? Recall first that one of our theoretical points of departure in this paper was the prediction that variational differences between styles and, by extension, registers are “quantitatively simple” (Guy 2005: 562; see also e.g., Labov 2010: 265; Rickford 2014: 596). In other words, probabilistic grammars should be fairly identical across registers. Based on our data we can assert fairly confidently that in terms of the probabilistic conditioning of grammatical variation, differences between registers are not quantitatively simple: registers do differ in terms of probabilistic complexity.

### Table 9
The dative alternation in Dutch: number of constraints (according to regression analysis), number of interactions (according to conditional inference tree analysis), and relative importance of lexical conditioning (according to regression analysis). The overall complexity ranking (rightmost column) is determined by calculating the arithmetic mean of the complexity ranks according to individual criteria. Lower-numbered ranks indicate more complexity.

<table>
<thead>
<tr>
<th></th>
<th>Number of constraints</th>
<th>Number of interactions</th>
<th>Relative importance of lexical conditioning (recipient)</th>
<th>Relative importance of lexical conditioning (theme)</th>
<th>Rank aggregation (mean rank)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spoken informal</td>
<td>5 (rank 3)</td>
<td>3 (rank 4)</td>
<td>0.48 (rank 3)</td>
<td>2.44 (rank 3)</td>
<td>3.25</td>
</tr>
<tr>
<td>Spoken formal</td>
<td>6 (rank 2)</td>
<td>6 (rank 2)</td>
<td>2.89 (rank 1)</td>
<td>1.59 (rank 4)</td>
<td>2.25</td>
</tr>
<tr>
<td>Written informal</td>
<td>5 (rank 3)</td>
<td>5 (rank 3)</td>
<td>0.70 (rank 2)</td>
<td>3.75 (rank 1)</td>
<td>2.25</td>
</tr>
<tr>
<td>Written formal</td>
<td>7 (rank 1)</td>
<td>7 (rank 1)</td>
<td>0.33 (rank 4)</td>
<td>3.26 (rank 2)</td>
<td>2.00</td>
</tr>
</tbody>
</table>
Second, contra quantitative simplicity claims one could have plausibly hypothesized that the complexity of particular registers should be proportional to the extent to which (a) particular registers are harder to master than others, (b) production is unconstrained by on-line processing limitations, and (c) production is subject to normative pressures. And these hypotheses are more in line with our results. A case in point is our finding that formal registers are consistently the most complex ones in the case studies investigated here. This pattern certainly ties in with gut feelings that many register analysts share about complexity variance. Consider face-to-face conversation: Biber et al. (2021b: 3) note how “most native speakers can fluently produce conversational discourse […] with no special effort at all, in real time with no planning, revision, or editing” – thus compared to a demonstrably formal register such as academic prose (which requires training to master and effort to produce), conversation would seem to be comparatively non-complex. We submit that the formal registers we cover in our study (parliamentary speeches and quality newspaper writing) are, like academic prose, characterized by acquisitional complexity and production complexity, implying that not all language users equally master and/or acquire them.

As to differences between spoken and written registers, our analysis shows that the written formal register (quality newspaper writing) tends to be (one of the) most complex registers under study here, while spoken registers are consistently the least complex ones. Why is that? We know that spoken registers “are produced and comprehended in real-time, setting a cognitive ceiling for the syntactic and lexical complexity typically found in these [registers]” (Biber 1988: 163). The on-line nature of spoken language (especially conversation) is why the production of spoken language is subject to processing and production constraints and biases (e.g., Hawkins 1994; MacDonald 2013) in a way that the production of written language is not. Monitoring is another important factor here: while speech (and especially informal, vernacular speech) is “the style in which the minimum attention is given to the monitoring of speech” (Labov 1972: 208), written language is more “governed by prescription” (D’Arcy and Tagliamonte 2015: 255). Prescriptivist pressures and the additional monitoring that comes with them, in turn presumably induce probabilistic complexities in written registers – and also, of course, in more formal registers.

6 Conclusion

In this paper, we have investigated the comparative probabilistic/variationist complexity of four registers at the intersection of formality and mode. We operationalized complexity as a function (a) of the extent to which variation is
regulated by more language-internal constraints, (b) of the extent to which constraints interact with each other, and (c) of the extent to which variation is lexically conditioned. Based on datasets covering grammatical alternations in two Germanic languages (English and Dutch), we demonstrated that formal registers are consistently the most complex ones, while spoken registers are the least complex ones. The most complex register under study is written-formal quality newspaper writing. On the interpretational plane, we have argued that complexity differentials are a function of acquisitional difficulty, of on-line processing limitations, and of normative pressures.

The methodology that we have adopted in this study is novel in that the vast majority of previous studies on the register-complexity nexus have adopted text-linguistic methodologies in the spirit of Biber (1988). This line of research has yielded a large body of invaluable results, some of which we reviewed in the Introduction section. But with that being said, what we did in this study is to focus not on texts or text frequencies, but on choice contexts and on the probabilistic conditioning of grammatical variation: when speakers and writers have a choice between different ways of expressing the same meaning or function, how complex is the choice-making process? To be clear, we believe that this is a complementary – not competing – perspective which is, however, well worth exploring more in future research. According to our results, register differences and complexity differentials between registers are not only about text frequencies and co-occurrence patterns of linguistic features. Rather, register also shapes the arguably deeper probabilistic conditioning of linguistic variation. In other words, register turns out to be an even more fundamental determinant of linguistic variation and of language complexity than has been traditionally recognized.

Limitations of this study and directions for future research include the following. For one thing, our comparatively coarse operationalization of register may have failed to capture finer-grained complexity differentials within individual registers. What is more, the registers we have selected for investigation in this paper may also vary in terms of communicative purposes, a fact that may be confounded with the concept of formality that we rely on (see e.g., Biber and Egbert 2018). Future research should investigate these issues in more detail. Second, more research examining other alternations (beyond the dative alternation and the future temporal reference alternation) is required for the sake of obtaining a fuller picture about register-induced complexity variation. Last but not least, the investigation needs to be extended to other languages beyond English and Dutch. This kind of analysis will open up exciting avenues for new research agendas at the intersection of variationist sociolinguistics, register analysis, and cross-linguistic typology.
Acknowledgments: The authors are grateful to two anonymous reviewers for constructive feedback. The usual disclaimers apply. Funding by the Research Foundation – Flanders (FWO) (grant #G0D4618) is gratefully acknowledged. This paper is dedicated to Douglas Biber, a great scholar whose work and kindness has inspired many, including the authors of the present paper.

References


