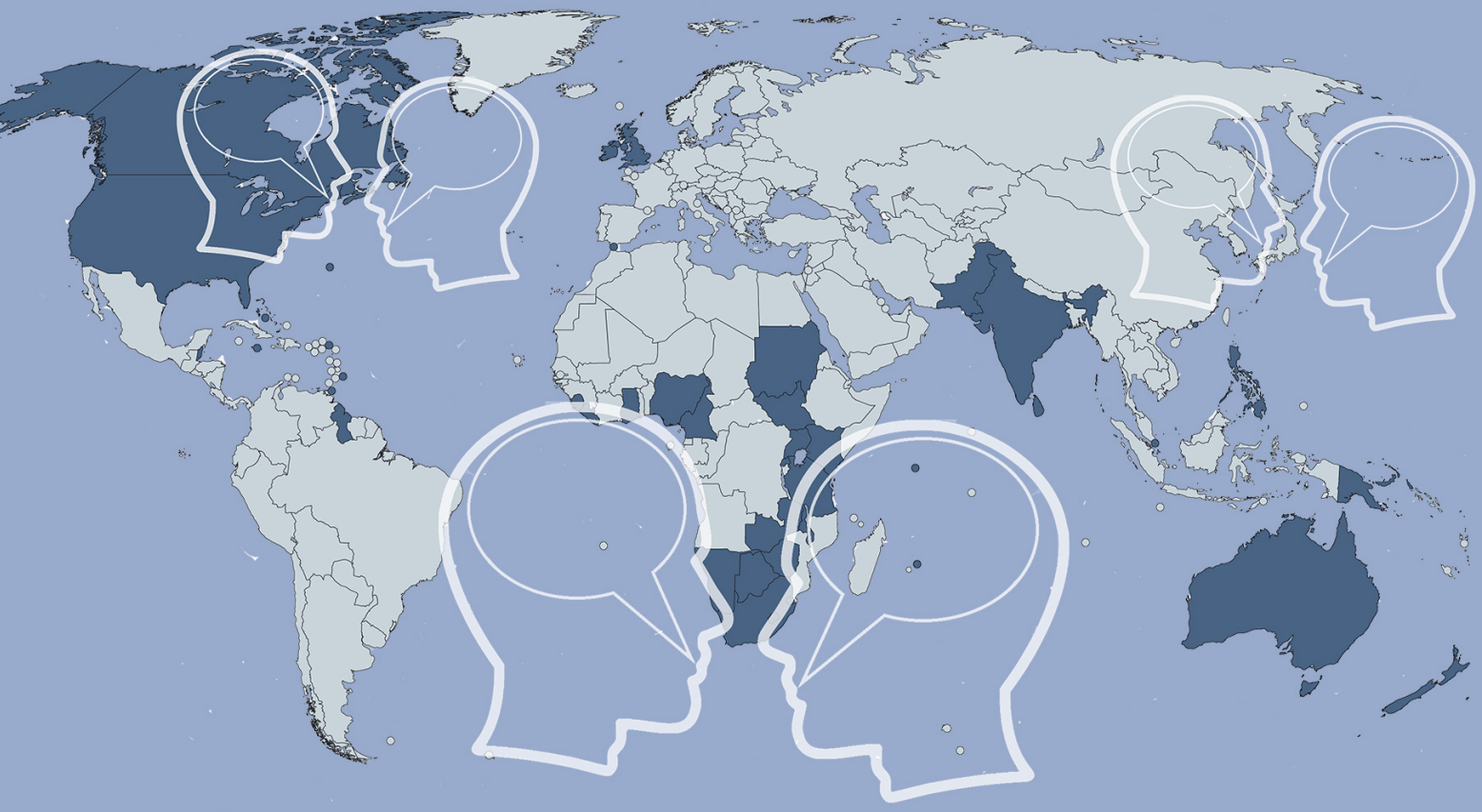
Regional variation in probabilistic grammars

A multifactorial study of the English dative alternation



Melanie Röthlisberger

2018

KU Leuven
Faculty of Arts
Department of Linguistics

Regional variation in probabilistic grammars

A multifactorial study of the English dative alternation

Dissertation presented in partial fulfillment of
the requirements
for the degree of Doctor in Linguistics

by Melanie Röthlisberger

Supervisor:	Prof Dr Benedikt Szmrecsanyi
Co-supervisors:	Dr Jason Grafmiller
	Prof Dr Marianne Hundt

Leuven 2018

Acknowledgements

This thesis would not have been possible without the tremendous support of numerous people. First and foremost, I am extremely grateful to my supervisors at KU Leuven, Benedikt and Jason, for their encouragement and valuable feedback, for always having '5 minutes', for the inspiring discussions and for their unwavering support the last four years. I would also like to thank Dirk Geeraerts, Marianne Hundt, Hubert Cuyckens, Karen Lahousse and Lieselotte Anderwald for accepting the invitation to act as member of my examination committee, and Liesbet Heyvaert for acting as the Chair of the committee. Marianne Hundt, Dirk Geeraerts and Hubert Cuyckens have additionally been members of my supervisory committee, and I wish to thank them for the interest they have shown in my work and for their support when I needed help, be it with feedback on my writing or providing me with reference letters for funding acquisition. Special thanks also go to Benedikt Heller (Benedikt "the Younger") – my co-PhD fellow – who always had an open ear for my questions related to perl and R and who (let's be honest) was the first to show me the vegetarian side of life. I also experienced the full advantages of being part of such an unbelievably amazing research group like QLVL. A big thank you goes to my fellow PhD colleagues: Alek, Danqing, Dirk P., Isabeau, Jocelyne, Karlien, Laura, Leonie, Mariana, Milly, Robbert, Stefano and Thomas – thank you so much for all the fun and laughter and the best coffee breaks one could think off! I am also indebted to the other senior members of QLVL, Dirk S., Eline, Freek, Kris, Kristina and Stefania, for providing such a supportive research atmosphere. An additional thank you goes to Eline, who types as rigorously as she speaks and is always up for crazy jokes. You are the best office mate (*ever*)! Laura, I cannot thank you enough for your friendship the past four years. Without you, I surely would not have made it through this enterprise the way I did. The supportive atmosphere of QLVL would not be the same were it not for Sylvia and Sonja who jointly master the administrative side of things and make the department run so smoothly. Thank you both for always being there for questions and for your support.

Apart from the cozy QLVL group, there were not many people I met in the past four years but the ones I met were extraordinary. Lena, you are my power girl! Thank you for your inspiration to put on more weight, for the encouraging text messages, Grijs nights and for always having a smile on your face when we were sweating our you-know-what off. A huge thank you also goes to Jack, Katharina, Lex, Naomi, Sali and the rest of the Linguistics department crew for making me feel welcome at UofT, for providing me with insights into variationist methods I would otherwise never have acquired and for giving me another home away from home away from home. I would also like to thank people outside of Leuven and Toronto who supported me those past years: Zeltia, who was my conference buddy all those years ago and my colleagues in Zurich who were there for me from the very beginning of my academic life.

There have also been a lot of people outside the academic world who are just dimly aware of what I have been doing and who gave me their friendship and love those past years. To all my scouting friends, Antares plus, Dunants and Schöftlers: M-E-R-C-I for keeping me firmly in reality when my mind was riding rollercoaster somewhere in the ivory tower. I also cannot thank my family enough for their love, their time, the good food and for my little godson. My biggest gratitude goes to Claudio who supports me every step of my life, who listens to my stories (more than once) and who always challenges me to give my very best and try out new things. Finally, I would also like to thank Christy Ha for her thorough and professional proofreading of my thesis. All remaining errors are of course my own.

Leuven, December 2017

Table of Contents

List of Tables

List of Figures

1

Introduction

> There must be the so-called theme and recipient, such as *John gives Mary the apple*.
>
> *<ICE-HK:W1A-020#90>*

Situated at the crossroads of variationist linguistics, cognitive sociolinguistics, and the probabilistic grammar framework, this study is among the first to provide a comprehensive description of the cross-lectal variability of probabilistic constraints that fuel syntactic variation in postcolonial varieties of English. To this end, the present work investigates the effect of conditioning factors that influence the choice between two syntactic variants which form part of the well-known dative alternation – namely the ditransitive dative (as in 1a) and the prepositional dative (as in 1b) – across nine national varieties of English.

(1) a. the ditransitive dative variant
So you also give [the police]$_{recipient}$ [a statement]$_{theme}$ <ICE-JA:S1B-069:1:A>[1]

b. the prepositional dative variant
You gave [the statement]$_{theme}$ to [the police]$_{recipient}$ <ICE-JA:S1B-069:1:A>

[1] Source labels include information on the corpus (ICE vs. GloWbE), variety (here: JA = Jamaican English), the genre (here: S1B), text number and possibly information about speaker (here: A); see also Section 4.3 on the corpus metadata.

Abundant research has shown that the factors governing the alternation between these two variants are multifaceted and non-deterministic: No single factor (or set of factors) categorically determines the choice of a given variant (see, for instance, Bernaisch et al. 2014; Bresnan et al. 2007a; Theijssen et al. 2013). Instead, numerous factors probabilistically influence the variation between the prepositional and the ditransitive dative. These factors include, for instance, pronominality, givenness, definiteness, frequency, animacy and length of the respective constituents (e.g. pronominal recipients favour the ditransitive, pronominal themes the prepositional dative), along with the semantics of the token in question: Abstract uses of *give* – *give them a break* – favour the ditransitive, while uses representing physical transfer – *give my card to them* – favour the prepositional dative variant. From a comparative perspective, there is some evidence that these factors may vary in subtle ways across different speech communities (Bresnan & Ford 2010; Tagliamonte 2014); however, the extent of conditioning factors' potential cross-lectal variability is still not well understood (see Bernaisch et al. 2014). By drawing on production data from nine national varieties of English covering 14 different registers and by including an extensive set of dative verbs, this large-scale comparative study contributes to patching the hole in our current understanding of the English dative alternation.

My primary interest thus lies in delimiting the scope of syntactic variation within and among different varieties of English around the world. In essence, I am interested in the extent to which speakers of different varieties of the same language rely on the same processes and/or probabilistic cues when choosing between dative variants.

I approach this overarching interest by means of a set of four more specific research questions:

- What is the extent to which varieties of English share, or do not share, a probabilistic grammar that is explanatory across different varieties? And what are the limits of cross-varietal variation?

- Are lectal differences random or can they be explained by considering socio-historical factors such as language contact?

- To what extent are factors that are typologically robust cross-lectally variable?

- Which of the individual constraints are tied to stylistic differences or lexical considerations?

To address these research questions, the study explores variability in the probabilistic constraints that fuel variation within and across speech communities. Theoretically, I assume a model of grammar that is at its core dynamic, probabilistic and usage-based (e.g. Bybee & Hopper 2001), and extend this model to my investigation of cross-varietal syntactic variation à la Bresnan and Hay (2008) and Bresnan and Ford (2010). My study thus fits in squarely with the theoretical frameworks of Probabilistic Grammar, Variationist Linguistics and the emerging field of Cognitive Sociolinguistics. Taken together, these three frameworks provide a model of grammar that views variation in language as an inescapable consequence of human interaction and asserts that linguistic events, such as the choice between two dative variants, can only be understood systematically when the social, contextual as well as cognitive aspects of language usage are taken into account.

Probabilistic approaches to language assume that grammar is inherently variable, shaped by social, cognitive or functional factors that are gradient in nature, that influence linguistic choice making in subtle ways (e.g. Bod et al. 2003; Bresnan 2007) and which can, when aggregated, result in population-level linguistic phenomena (Scott-Phillips & Kirby 2010: 411). In this context, a speaker's probabilistic grammar constitutes this individual's probabilistic knowledge about the constraints that influence all aspects of language production and comprehension. One such probabilistic constraint is for instance the *Principle of end-weight* (Behaghel 1909), which posits that language users prefer constituents in the order of increasing size or complexity to ease processing and production. This principle is not necessarily tied to surface material but to more or less subtle stochastic generalisations about language usage, which – according to experimental evidence – language users implicitly know about (e.g. Bresnan 2007). These subtle stochastic generalisations are pervasive on all levels of language (see, for instance, the contributions in Bod et al. 2003).

Like all usage-based accounts, probabilistic models of grammar are committed to the notion that grammars are learnt from exposure to other speakers and are "the cognitive organization of one's experience with language" (Bybee 2006: 711). To the extent that the linguistic experience of different speakers and communities varies, successive generations of speakers will give rise to variation in language as every new generation adapts their own linguistic knowledge to match that of their input. From that perspective, we expect differences between speech communities to emerge in those contexts where shifting usage frequencies have led to changes in the probabilistic constraints that shape linguistic variation. As Bresnan & Hay (2008:

246) put it:

> A probabilistic, usage-based approach to grammar is able to account for such variation by assuming that different communities differ in the types and frequencies of the constructions that they are exposed to. However, a probabilistic approach also predicts that variation across space and time should exist in less obvious ways – even affecting the subtle probabilistic choices that are made between two variants which are equally acceptable for that dialect. That is, we expect to observe syntactic differences in time and space which are reflected not only in the use of clear dialectal features or clear-cut changes in progress, but also in extremely subtle factors such as the relative probabilistic weights of conditioning factors, and changes over time in speakers' preferences between equally well-formed variants. (Bresnan & Hay 2008: 246)

Szmrecsanyi et al. (2016) call this process *probabilistic indigenisation* and describe it as the gradient localised acculturation of probabilistic constraints due to shifting frequency patterns in language internal variation. The "emergence of locally characteristic linguistic patterns" in new varieties of English (Schneider 2007: 6) hence does not only take place at the lexis-syntax interface but also on more fine-grained levels of linguistic knowledge, namely in the underlying stochastic patterns that make up speakers' probabilistic grammar. Extending the terminology, Röthlisberger et al. (2017) stress the outcome of probabilistic indigenisation by referring to it as *cognitive indigenisation*, that is, the lectalisation or creation of distinct lects[2] at the level of very subtle gradience (Röthlisberger et al. 2017: 677).

Importantly, the process referred to as probabilistic indigenisation ties in probabilistic approaches to grammar with the concept of indigenisation in research on World Englishes. To model language users' probabilistic knowledge across regionally distinct varieties of English, the present study will thus turn to the variation-centred, usage- and experience-based Probabilistic Grammar framework developed by Joan Bresnan and colleagues (Bresnan 2007; Bresnan et al. 2007a; Bresnan & Hay 2008; Bresnan & Ford 2010). This framework relies on two crucial assumptions:

- Grammatical variation is sensitive to multiple and typically conflicting probabilistic constraints, be they formal, semantic, or contextual in nature. Such

[2]Lect is an umbrella term used in Cognitive Sociolinguistics to refer to the collection of linguistic features that vary along external contextual dimensions such as region or social class. We can thus distinguish, for instance, regiolects, dialects, ethnolects, sociolects, idiolects, etc. (see Geeraerts et al. 1994: 4; Geeraerts 2005)

constraints, like the principle of end-weight, may influence linguistic choice-making in subtle ways.

- Grammatical knowledge must have a probabilistic component, for the likelihood of finding a particular linguistic variant in a particular context in a corpus has been shown to correspond to the intuitions that speakers have about the acceptability of that particular variant.

Research traditions and methodologies developed within a probabilistic approach to grammar are by and large compatible with work in Variationist Sociolinguistics (see Labov 1982). Common to both research fields is their interest in linguistic variation and in the constraints that influence speakers' choice between "alternate ways of saying 'the same' thing" (Labov 1972a: 188). Indeed, variationist sociolinguists traditionally take a quantitative perspective on inter- and intra-systemic variation and have long made use of naturalistic corpus data (such as collections of sociolinguistic interviews) to, for instance, compare variation patterns between speaker groups or across communities (e.g. Tagliamonte 2002). Probabilistic approaches are also consonant with research in the developing subfield of Cognitive Sociolinguistics which aims to integrate social meaning and socially conditioned variation with a cognitive dimension to offer a more complete model of language structure and variation (Geeraerts et al. 2010; Pütz et al. 2014). So what do Variationist Sociolinguistics and Cognitive Sociolinguistics have to offer to the current endeavour that a probabilistic account by itself cannot provide?

Variationist Sociolinguistics as a field of research took off with William Labov's study of speech patterns in Martha's Vineyard and New York City (Labov 1963, 1966, 1972a; see also Weinreich et al. 1968). Labov's work soon inspired others to apply a variationist approach to the analysis of linguistic patterns in a wide variety of communities around the world, including Panama (Cedergren 1973), Norwich (Trudgill 1974), Anniston (Feagin 1979) and Guyana (Rickford 1987) (Bailey 2013: 117; see similarly Milroy 1980 and Fasold 1984). Research within a variationist account assumes systematic and inherent variation in language, so-called structured heterogeneity, which can be analysed quantitatively (Labov 1972a). Due to this interest in quantitative analysis, Variationist Sociolinguistics was one of the first fields in linguistics that invoked statistical techniques to explore the structured heterogeneity in language (see Cedergren & Sankoff 1974; Sankoff & Labov 1979; Sankoff 1988; Tagliamonte 2012).

The main task of a variationist sociolinguist is to spot variable structures in language and to correlate this linguistic variation with some linguistic and non-linguistic parameters related to the grammatical context, the social context, community settings or registers, among others (see Chambers 2009: 18; Tagliamonte 2012: 7). These variable patterns are then interpreted by employing statistical techniques to determine the statistically significant factors on speakers' choice, their contribution and their relative importance (Tagliamonte 2012: 7). All studies within the framework of variationist sociolinguistics strive to adhere to the *Principle of accountability*: Linguistic variation must be studied in the context of the subsystem of which it is part. Hence, when analysing the linguistic pattern of a variable in its respective context, competing variants within the same context must also be considered.

One main advantage of Variationist Sociolinguistics is the rigorous methodology applied to the probabilistic analysis of variation patterns. Such probabilistic analyses are especially imperative for the Comparative Sociolinguistic method – a subfield of Variationist Sociolinguistics – which seeks to compare community grammars inter- and intra-systemically (Tagliamonte 2002). Similarities between varieties or lects with regard to the statistical significance of constraints, their relative importance and the constraints' effect sizes are taken as an indication of a common source of shared dialect features. According to Tagliamonte (2002: 731-733), researchers should adopt the following procedure in order to appropriately compare lects in a comparative sociolinguistic fashion:

1. Select an appropriate linguistic feature,

2. examine the patterns of use and define the conditioning constraints that contribute to variation within that linguistic feature, their statistical significance, relative strength and their ranking within one variety and

3. compare and contrast conditioning factors across sets of data with regard to

    - statistical significance,
    - relative strength and
    - constraint hierarchy.

The constraints on variation and their ranking provide two critical measures of comparison between varieties and dialects: If the conditioning factors and their ranking

are shared across a set of varieties, we can infer that they have inherited them from a common source (Tagliamonte 2002: 731).

Similarly to variationist and comparative sociolinguistic research, the present work seeks to compare the probabilistic grammar(s) of varieties by assessing patterns of usage, the circumstances of variation and by identifying the possible causes of variation. The conditioning factors' statistical significance, their relative strength and their hierarchical ranking thereby take centre stage in the analysis. Comparative sociolinguistic methods are especially well suited to explore regional differences between varieties of English. Taking the founding hypothesis of comparative sociolinguistics one step further, we can even hypothesise that the less two varieties diverge with regard to the structure of their conditioning factors, the more recent their separation and/or the more they share a common sociolinguistic reality.

Finally, the current study also shares its interest in the social and cognitive aspects of variation with recent research in Cognitive Sociolinguistics – a subdiscipline of Cognitive Linguistics that merges the main viewpoint of Cognitive Linguistics, namely that language is entrenched within one's general cognitive abilities, with a sociovariationist view, that is, an interest in the social and cultural forces that drive variation in human interaction (see, e.g. Geeraerts et al. 2010; Harder 2010; Kristiansen & Geeraerts 2013). Cognitive Sociolinguistics acknowledges the inherent heterogeneity of language as a social construct and is concerned with the effect that cognitive and sociocultural forces exert on the formation of distinct lects. From a cognitive (socio-)linguistic perspective, variationist studies such as the present one can be seen as investigations of the forces shaping the interaction between "formal onomasiological variation" and "speaker and situation related variation" (Geeraerts et al. 2010: 7-8).

The present study also ties in with psycholinguistic approaches to grammatical structure and variation, as in MacDonald (2013), which assume that language users are subject to the same psychological processes shaping production and comprehension and are thus likely to make similar syntactic choices, all else being equal. For instance, Bresnan et al. (2007a) find that in conversational American English, speakers are more likely to choose that dative variant which places the 'easier' or more accessible constituent before the less accessible one. Their findings are consistent with MacDonald's (2013) *Easy First* principle which refers to the general bias of language users to place 'easy' elements first in utterances. 'Easy' in this sense designates those elements that are more quickly retrieved from (long-term) memory (MacDonald 2013:

4), and an element may be easier to retrieve by virtue of it being more frequent, shorter, less syntactically complex, more conceptually salient or having been recently mentioned. Uttering the easier elements first gives the speaker enough time to plan and produce the more difficult constituents. The effects of the various factors constraining dative choice are coherent in that regard as animate, given, pronominal, definite and short constituents are all 'easy' elements. Such a harmonic alignment in the effect of the probabilistic constraints shaping variation has been observed in numerous native and non-native varieties of English (e.g. Bresnan & Hay 2008; Bresnan & Ford 2010; De Cuypere & Verbeke 2013; Bernaisch et al. 2014; Tagliamonte 2014). Bernaisch et al. (2014), for example, who analyse the dative alternation with *give* in South Asian varieties of English, argue that the factors determining the choice between the two dative variants can be universally applied to all varieties of English and are independent of the regional background of language users. Despite the persistent evidence that these general statistical tendencies and processing principles underlying the dative alternation are shared across a large set of varieties, other studies have found subtle effects of probabilistic indigenisation. Recent work demonstrates that syntactic choices within and across varieties are governed by language-internal forces that can exhibit subtle degrees of variability across regions (e.g. Bresnan & Hay 2008; Mukherjee & Hoffmann 2006), time (e.g. Wolk et al. 2013) and register (e.g. Gries 2013; Grafmiller 2014).

The current work is thus a symbiotic one. Not only do all of these theoretical accounts contribute jointly to a better understanding of the underpinnings of syntactic variation in World Englishes, the present study itself advances theory formation in Variationist and Cognitive Sociolinguistics by offering a detailed account of the underlying constraints that govern syntactic variation from a large-scale comparative perspective. This large-scale comparative perspective incorporates the cognitive as well as the social dimension of linguistic variation to account for the effects of probabilistic indigenisation thereby also offering novel insights into the indigenisation process itself.

The present work is structured as follows: Chapter 2 introduces the extensive body of previous research on the English dative alternation in two main parts. The first part presents a brief diachronic account of the English dative alternation from Old English to Present-Day English and traces the development of the two dative variants over time. The second part introduces studies that have analysed the dative alternation from a synchronic perspective from generativism to research in World Englishes. Since

the main focus of the current study is on cognitive-functional, corpus-based accounts of the dative alternation, the insights of other research traditions will only briefly be discussed. The trajectory of the described synchronic research will highlight the necessity of the present work.

Chapter 3 provides the socio-historical context for this study's analysis of regional variation. The chapter starts by introducing three models that have categorised varieties of English based on type, historical and evolutionary development. The chapter then focuses separately on each of the nine varieties under scrutiny in the present work, sketches their historical and linguistic background and situates them within the three aforementioned models. Nine varieties take centre stage in the analysis, namely British English (BrE), Canadian English (CanE), Hong Kong English (HKE), Indian English (IndE), Irish English (IrE), Jamaican English (JamE), New Zealand English (NZE), Philippine English (PhiE) and Singapore English (SinE). As will be shown, each variety is fairly unique in its socio-historical setting and can only be situated on an aggregate level within the three proposed models. Nevertheless, these models serve as a basis from which the present work can draw useful generalisations in its investigation of regional variation.

Chapter 4 presents the data and methodology and describes the statistical toolkit used for the analyses. Dative observations were drawn from both the *International Corpus of English* and the *Corpus of Global web-based English*. Non-alternating variants were excluded and the remaining $N = 13{,}171$ observations were annotated for numerous probabilistic constraints given the literature. The descriptive statistics provided for each constraint indicate a general tendency of language users to place 'easy' elements first. The statistical toolkit introduces random forest analysis, mixed-effects logistic regression and dialectometric techniques (i.e. distance metrics and multidimensional scaling) and highlights potential shortcomings and advantages of each for the subsequent analyses.

Chapter 5 presents the results of the various analyses one by one. Using mixed-effects logistic regression and random forest techniques, I will show that relative length of constituents and recipient pronominality are not only the two most important predictors, as evidenced by the random forest, but also the two constraints – together with CORPUS – amenable to regional variation. Further probing into these three constraints reveals that other end-weight related factors are not as regionally variable as length even when using a fine-grained measurement and that the cross-varietal malleability of recipient pronominality can be linked to differences in the

lexical profiles of ditransitive and prepositional variants across varieties. A detailed examination of corpus effects, which boil down to register effects, reveals that native and non-native varieties are distinct from each other regarding the effect of register and that recipient pronominality is the predictor most amenable to cross-register differences. Finally, calculating the probabilistic distance between varieties along three dimensions provided by comparative sociolinguistic methods results in variety clusters that mainly pit American-influenced (CanE, PhiE) versus non-American-influenced varieties. All in all, the findings presented in this chapter highlight the extensiveness of the cross-lectal variability of probabilistic constraints suggesting that probabilistic grammar(s) might not be as stable as hitherto assumed.

The results of Chapter 5 are discussed in Chapter 6, which places the study's results within the explanatory framework of Cognitive Sociolinguistics and general biases in language production and planning. Furthermore, Chapter 6 contrasts the current study's findings with previous research and highlights the innovative aspect as well as potential challenges of the present work. As will be shown, the aggregate perspective adopted here is unprecedented in earlier work and enables a detailed and comprehensive investigation of regional variation in probabilistic grammars worldwide. What is more, the observed ubiquity of variation in probabilistic grammars puts our understanding of lectal stability to the test.

The study ends with Chapter 7 which offers concluding remarks on the stability of probabilistic grammars, on the theoretical implications of the current work for research in Variationist Sociolinguistics, Cognitive Sociolinguistics and studies on World Englishes and sketches directions for future research.

With the exception of Chapter 4 (Methodology), Chapter 6 (Discussion) and Chapter 7 (Conclusion), each chapter concludes with a short summary.

2

The dative alternation

2.1 Introduction

Most basically, dative constructions involve a ditransitive verb that takes two semantic roles, namely a recipient-like and a theme-like argument (see Malchukov et al. 2010: 1). Similar to other Germanic languages, English ditransitive verbs typically occur in or alternate between a nominal and a prepositional pattern. The term *dative alternation* is thereby used to refer to the variation between these two patterns in Standard English, that is, the variation between the ditransitive dative (i.e. the nominal pattern in 2a) and the prepositional dative in (2b). While more dative constructions are theoretically possible (e.g. *John gives the apple Mary*) – prominently so in British English dialects (see Gast 2007; Siewierska & Hollmann 2007; Gerwin 2014) – these additional, rather infrequent, patterns are not of concern for the current study.

(2) a. the ditransitive dative variant
So you also give [the police]$_{recipient}$ [a statement]$_{theme}$ <ICE-JA:S1B-069:1:A>

b. the prepositional dative variant
You gave [the statement]$_{theme}$ to [the police]$_{recipient}$ <ICE-JA:S1B-069:1:A>

The objective of the present chapter is to provide an overview of earlier work on the English dative alternation in order to illustrate that the current study constituted the inevitable next step. The chapter is structured as follows: Section 2.2 sketches

previous work that analyses variation and change in the English dative alternation from a diachronic perspective, thus offering insights into the history of the English ditransitive and prepositional dative. Sections 2.3 to 2.6 introduce synchronic work on the English dative alternation from early generative accounts to usage-based research in World Englishes. The vast amount of research on the English dative alternation is ample testimony that the dative alternation constitutes one of the best-researched syntactic alternations in English. Needless to say, the unsurmountable amount of scholarly work cannot all be discussed in the present chapter. Instead, this chapter will focus on research relevant for the present analysis and only briefly touch upon other accounts where deemed necessary. Research will be presented in chronological order of the moment a research tradition first took interest in the dative alternation, from early formal accounts, language acquisition research and variationist approaches to, finally, probabilistic grammars and World Englishes. A summary of the chapter is provided in Section 2.7.

2.2 Diachronic aspects of the dative alternation — A brief historical sketch

While synchronic descriptions of the English dative alternation have received ample attention in the literature (see, for instance, Bresnan et al. 2007a; Kendall et al. 2011; Theijssen 2012; Schilk et al. 2013, among others), the diachronic development of this alternation is less well documented (Wolk et al. 2013: 385) and has only recently started to be investigated more. Diachronic studies have thereby mainly concentrated on changes in the available patterns of ditransitive verbs, changes in the formal and functional features of the respective variants (such as the preferred order of objects and the factors influencing it or the range of verb classes associated with the patterns) as well as the role played by morphological case marking in these developments (e.g. Allen 1995; McFadden 2002; Colleman & De Clerck 2009; Barðdal et al. 2011; De Cuypere 2015a, 2015b; Zehentner 2016, 2017).

Results of these studies highlight that prepositional datives were already in use in Old English but lexically restricted to verbs of caused-motion and communication (Allen 2006: 206; De Cuypere 2015b: 2). That is, the preposition *to* did not radically replace the Old English dative case in Middle English (Allen 2006: 214) as has often been assumed. Rather, the corpus data suggests that there was a gradual increase in the prepositional option: In Old English, variants with and without preposition

(*John gives it to her* vs. *John gives it her*) were used side by side, with the latter being the more conservative (Gast 2007: 52) and the former the more innovative pattern (Gerwin 2013: 457). In the transition to Middle English, the prepositional dative variant slowly encroached on the turf of its paradigmatic (alternative) counterpart in the case of two nominal objects (e.g. *John gives the apple Mary*) and seemed to have become a full-fledged alternative to the ditransitive by the Middle English period (McFadden 2002: 112). By the late fourteenth century, the alternative pattern with two nominal objects, as in *John gives the apple Mary*, had disappeared (Allen 2006: 206). After the Middle English period, and especially between the late seventeenth and early eighteenth century, the prepositional variant with two pronominal objects (*John gives it to her*) increased in frequency relative to its non-prepositional alternative (*John gives it her*). And by the late twentieth century, the present-day prepositional pattern had ousted its alternative non-prepositional variant completely in Standard English (Yáñez-Bouza & Denison 2015: 255). Syntactic transfer from French (where the dative is always marked with a preposition), the ready availability of prepositional dative variants from Old English onwards and a general increase of analytic case marking must have played a role in the rapid upsurge of *to*-datives and the parallel loss of the alternative non-prepositional variant (see Allen 2006: 214). While the syntactic pattern without the preposition (e.g. *John gives the apple Mary*) has thus disappeared from Standard English, it remains part of dialectal grammars in the British Isles (see Gast 2007; Siewierska & Hollmann 2007; Gerwin 2014). Today, non-prepositional ditransitive patterns (both the dialectal and the standard one) are increasing in frequency in spoken twentieth-century British English at the expense of the prepositional dative variant. Nevertheless, the latter variant has retained its popularity in written language (Gerwin 2014: 201).

Besides this interest in the distributional patterns of the ditransitive and prepositional dative over time, work has also been under way to investigate the various constraints that fuel the variation between the two dative variants and these constraints' malleability over time. Studies by De Cuypere (2010) and Wolk et al. (2013) suggest that the probabilistic constraints influencing dative choice have by and large remained relatively stable diachronically. At the same time, the results of the two studies indicate that the strength of the effects of length and animacy has undergone significant changes in the course of time (Wolk et al. 2013: 405). Similarly, Bresnan & Hay (2008), who analyse diachronic changes in the constraints influencing dative choice in New Zealand English, observe that although the probabilistic grammar of

speakers is diachronically very robust, subtle but significant differences emerge in the strength of the factor 'length'.

All in all, earlier work with a diachronic perspective not only observed changes in the range of verb classes associated with the ditransitive and prepositional variant but also functional changes in the probabilistic constraints shaping the variation – particularly regarding the effect of constituent length.

The recent upsurge in diachronic studies notwithstanding, the majority of studies on the English dative alternation have generally taken a synchronic perspective to analyse the variants' variable patterns. For decades, linguists interested in the English dative alternation have thereby been split into two opposing camps: Those who adopt a *single-meaning approach* and those who adopt a *multiple-meaning approach*. Authors who adopt a single-meaning approach posit that both variants are essentially semantically equivalent; the choice between them is driven by language-internal factors pertaining to the verb and the two objects. The multiple-meaning approach asserts that because the two constructions are syntactically different, they are also different semantically and hence represent two different event structures (Gerwin 2014: 19).

2.3 Formal models of the dative alternation

The division between scholars presuming a single-meaning and those arguing for a multiple-meaning approach has carried through most linguistic inquiries into the dative alternation. In the heyday of generativism, linguists who followed the multiple-meaning approach distinguished between the two realisations of dative constructions based on verb semantics. Under the assumption that verbs that share the same syntactic behaviour are also semantically similar, researchers have aimed to group alternating and non-alternating verbs according to their semantics (e.g. Green 1974; Levin 1993). As a consequence of the variants' semantic differences, they then argued for differences in the deep structures of each dative variant (e.g. Green 1974; Oehrle 1976; Baker 1979). In contrast, the single-meaning approach in generativism assumed an identical structural relationship between the ditransitive and prepositional dative variant at the level of the deep structure whereby one variant was taken to be original and the other constructed via a transformational rule (e.g. Larson 1988).

The focus of generativism on deep structures and hard-wired, innate grammar was countered by the emergence of usage-based approaches in linguistics which

argue that speakers' grammars are formed from linguistic experience. As one of the formal approaches that soon ascribed to the usage-based perspective (see, for instance, Goldberg 1995: 7; Goldberg 2003: 222), Construction Grammar groups a number of models that all assume that grammar is made up of form-function pairings, so-called *constructions*. These constructions can range from fully idiomatic and lexically instantiated ones to completely abstract patterns at all linguistic levels. Construction Grammarians argue that constructions as form-meaning pairings carry meaning themselves. The availability of different constructions, as in the case of the two dative variants, necessarily entails different meanings associated with them. Similar to generative approaches, the meaning of a dative variant was inferred from the meaning of the specific verbs used (Goldberg 1995). Note, however, that Construction Grammar posits non-compositionality of constructions' meaning. In other words, the meaning of a construction remains unpredictable from its components.

Constructions are seen as abstract syntactic patterns with empty slots that can be filled by concrete (lexically instantiated) words or other (abstract) constructions. The combination of various constructions create a hierarchical network, the *construcicon*. Only on the lowest level of abstraction can we find concrete instantiations of constructions, so-called *constructs*, as in *John gives Mary the apple*. Even though the prepositional dative and ditransitive dative variant might be semantically linked on some extended abstract level (both relating to a metaphor called TRANSFER OF OWNERSHIP AS PHYSICAL TRANSFER), they are not syntactically synonymous (Goldberg 1995: 91). According to Goldberg's *Principle of No Synonymy*, and because the two constructions are not motivated by each other (that is, hierarchically related constructions, see Goldberg 1995: 72), they have to be pragmatically different (Goldberg 1995: 67). Postulating pragmatic differences is nothing new. In fact, following the multiple-meaning approach advocated by Goldberg and others before her, some studies have taken a step beyond verb semantics and instead have started to focus on the pragmatic aspects that distinguish the two variants (such as information status, definiteness of the constituents, and so on).

2.4 First and second language acquisition research

Similar to Construction Grammar, earlier work in language acquisition also followed the generative tenor (e.g. Gropen et al. 1989) but soon reoriented itself to the usage-based perspective (e.g. Tomasello 2003). In first language acquisition, Campbell &

Tomasello (2001) show that the ditransitive dative is generally acquired before the prepositional dative (due to frequency of exposure) and that verbs are often used in that variant in which children were exposed to it first (Campbell & Tomasello 2001: 257, 266). These conservative tendencies have been observed by Gropen et al. (1989: 239), Dodson & Tomasello (1998: 617) and by Childers & Tomasello (2001: 743) in first language acquisition, and by Gries & Wulff (2005: 196) in second language acquisition. What is more, studies in first language acquisition which follow a probabilistic approach show that children's and adults' grammars are constrained by the same underlying factors when choosing between two dative variants. Children only differ from adults in the degree of their sensitivity towards these effects, that is, in their production probabilities (de Marneffe et al. 2012: 53; van den Bosch & Bresnan 2015: 110; see also Bürkle 2015).

The acquired production probabilities from a speaker's first language (L1) can interfere when acquiring a second language (L2) as the cue strength from their L1 is transferred to their L2 (see MacWhinney 1997). This interference loses in strength over time when speakers' L2 probabilistic grammar shifts to native-like settings (MacWhinney 1997: 129; Ellis 2006: 169). Differences between L1-like and L2-like uses of variants have been the main focus of corpus-related studies on the dative alternation in English as a Second Language (ESL) and English as a Foreign Language (EFL). For instance, Gries & Deshors (2015) explore deviations between native- and non-native speaker choices of dative variants using data from Learner Englishes, indigenised varieties of English and British English (as the native-like reference variety). They make use of a novel approach that involves mixed-effects logistic regression (see Gries & Deshors 2015: 139 for details). The results of their study indicate that non-native speakers make native-like choices in almost all contexts provided the cues (e.g. length, pronominality of recipient and theme) are strong enough. If cues are unreliable, non-native speakers tend to opt for the prepositional dative (Gries & Deshors 2015: 152).

Comparing first and second language acquisition research highlights three similarities in the acquisition process of the dative alternation: First, the production of ditransitive datives by language learners seems to be restricted to specific lexical items (Savage et al. 2003: 564; McDonough 2006: 194). Second, the ditransitive dative seems to be acquired first with pronouns and only later with fully lexicalised noun phrases (NPs) (Dodson & Tomasello 1998: 614; Childers & Tomasello 2001: 743; McDonough 2006: 194). And third, the prepositional dative variant is apparently the

preferred option in both first and second language acquisition (Conwell & Demuth 2007: 177; Jäschke & Plag 2016). De Cuypere et al. (2014) explain this third finding with Pieneman's *Processability Theory* (Pienemann 1998). They argue that the prepositional dative variant constitutes the more transparent option due to a direct mapping of the relationship between thematic roles, grammatical functions and constituents. As a result of that, the prepositional variant is easier to process and therefore the preferred option for language learners (De Cuypere et al. 2014: 193, 203).

2.5 Information, processing and variationist accounts

With the availability of large text collections starting in the 1990s and the increase in computational power, quantitative analyses of the contextual and psycholinguistic constraints on dative choice gained ground. Thompson (1990) was among the first to use corpus data to assess the influence of length, pronominality, identifiability, specificity, animacy and status of the recipient on dative choice (although the number of observations analysed was fairly small). A few years later, Williams (1994) applied parametric multiple regression in SAS to explore the multivariate nature and simultaneous influence of various constraints on the dative alternation (see Williams 1994: 44). His study is not only groundbreaking with respect to the statistical techniques used, Williams also includes previously neglected constraints on constituent ordering in his analysis, namely register, modality, syntactic class of the verb and prosodic length of the constituents. In a similar vein, Collins (1995) highlights the importance of accessibility, end-weight, pronominality and definiteness in the choice between prepositional and ditransitive datives using corpus data from Australian English. The results of his study indicate that the difference in communicative status between recipient and theme is stronger in the ditransitive than in the prepositional dative. Adopting a similar multivariate perspective, Arnold et al. (2000) draw on corpus and experimental data and show that grammatical complexity and discourse status influence the choice of dative variant simultaneously and partly independently from each other.

Studies interested in the pragmatic context of the choice between the dative variants paid attention to discourse constraints such as information status (e.g. Erteschik-Shir 1979; Thompson 1990; Collins 1995), the constituents' definiteness (e.g. Erteschik-Shir 1979; Collins 1995), their animacy status (e.g. Ransom 1979), their pronominality (see Collins 1995: 39; Aissen 2003: 437) and other lexical char-

acteristics (e.g. Wolfe-Quintero 1993). Other, more psycholinguistic-oriented work explained the ordering of constituents in the dative alternation by referring to such concepts as accessibility, processing demands and persistence (e.g. Smyth et al. 1979; Bock & Irwin 1980; Bock 1986; Bock & Griffin 2000; Gries 2005). The concepts of processing demands and persistence especially have been at the centre of research (e.g. Bock 1986; Hawkins 1994; Wasow 1997b; Rohdenburg 2002; Hawkins 2004; Szmrecsanyi 2005; Stallings & MacDonald 2011). Processing demands closely relate to end-weight effects and the tendency of language users to place more accessible items before less accessible ones (at least in English). End-weight effects, on the other hand, are so persistent in language that they have been argued to be solely responsible for all constituent ordering (Hawkins 1994) – a proposition that has been criticised in Wasow (1997a) and Wasow (1997b). Structural persistence (also known as syntactic priming) refers to speakers' tendency to reuse syntactic constructions. The ditransitive dative variant is thus more likely if the preceding variant in discourse was also a ditransitive dative, and the prepositional dative variant is more likely if the preceding variant was a prepositional dative (see Branigan et al. 2000; Gries & Wulff 2005; McDonough 2006). Corpus-based work highlights that results obtained from observational aggregate (corpus) data match results obtained from behavioural individual (experimental) data (Gries 2005: 387). What is more, results in Gries (2005) suggest that priming effects in the dative alternation are verb-specific.

This boost of studies using multifactorial methods moved the research focus away from the discussion of the single-meaning vs. multiple-meaning perspective prevalent in the later half of the twentieth century. Inspired by work in variationist sociolinguistics, (corpus) linguists turned their attention instead to the factors that drive the alternation of constructional variants in a carefully predefined envelope of variation (see Williams 1994: 37). They encountered one problem, however. Variationist sociolinguistics had started its endeavour analysing the factors that drive the choice of a sociolinguistic variable in the phonological context, for instance between the two (phonological) variants [ɪn] vs. [ɪŋ] to express *-ing* (see, e.g. Labov 1966). When linguists began to transfer phonological variation (where the variants might carry social but no propositional meaning) to lexical and grammatical elements (which carry meaning by definition) (e.g. Weiner & Labov 1983; Harris 1984; Sankoff et al. 1997), it became apparent that propositional equivalence in syntactic variation could not be defined in a straightforward fashion (see Lavandera 1978; Sankoff 1988). The most useful solution to the problem was to define a choice context where

two (or more) syntactic variants were equal under the same truth conditions. The focus was thus not so much on the semantic equivalence between variants but rather on "the study of the internal linguistic factors (e.g. syntactic, semantic, discourse/ pragmatic factors) which may influence the choice of a variant in a given context" (Silva-Corvalán 1986: 121). These internal linguistic factors influence the choice of a variant probabilistically and not categorically, as variationist sociolinguistic studies and probabilistic accounts show.

2.6 Probabilistic grammars and World Englishes

Variation-centred, usage-based probabilistic approaches to the English dative alternation took off with Bresnan et al. (2007a) who were the first to analyse comprehensively the simultaneous effect of conditioning factors on the choice of dative variant in American English while also emphasising the benefits gained from employing multifactorial techniques (see also Gries 2001). Using naturalistic production data, Bresnan et al. (2007a) highlight that the choice of dative variant is influenced by gradient constraints rather than absolute ones. Even allegedly idiomatic and fixed expressions at various degrees of lexicalisation (e.g. *drop sb. a line, give birth to sb.*) or light verb constructions (Elenbaas 2013) are shown to be variable between the two variants. The variation between the ditransitive and prepositional dative should arguably be regarded in terms of "pragmatic probabilities" (Bresnan & Nikitina 2003: 34) instead of categorical (un-)grammaticality (see Gerwin 2014: 54). Bresnan & Hay (2008) were among the first to illustrate that these pragmatic probabilities shaping variation in the dative alternation are also subject to social (that is, language-external) variation. Theirs and other studies within the probabilistic approach use state-of-the-art statistical techniques, such as mixed-effects logistic regression, random forests and conditional inference trees, to highlight that the probabilistic constraints on syntactic variation are malleable across time (Wolk et al. 2013), register (Grafmiller 2014) and space (Bresnan & Hay 2008; Grimm & Bresnan 2009; Bresnan & Ford 2010; Schilk et al. 2013; Bernaisch et al. 2014). The latter interest in regional variation in probabilistic constraints has especially taken centre stage in recent work. For instance, Bresnan & Hay (2008) compare the influence of constraints on dative choice between New Zealand and American English and observe that speakers of New Zealand English are more likely than speakers of American English to use a prepositional dative when the recipient is inanimate and a ditransitive dative when the recipient is animate.

This cross-varietal difference between New Zealand English and American English regarding the stronger effect of animacy in NZE seems to have been long-established since the nineteenth and twentieth century (Hundt & Szmrecsanyi 2012). Zooming in on the New Zealand data, Bresnan & Hay (2008) further illustrate that the oldest speakers (data from the 1850s) and youngest speakers (data from the 2000s) of New Zealand English display a quantitative preference for the prepositional dative while there is a drop in frequency around the 1900s (Bresnan & Hay 2008: 253-254). In a comparable study, Bresnan & Ford (2010) find probabilistic production differences between Australian and American speakers of English with regard to the effect of length on dative choice (2010: 169). They verify their corpus-based models with experimental judgment tasks and confirm the production probabilities obtained from their statistical models. The results of their study indicate that speakers of Australian English are more likely than their American counterparts to use a prepositional dative when relative length of themes increases (Bresnan & Ford 2010: 203). Similarly, Schilk et al. (2013) observe in their study on Southeast Asian Englishes that recipient pronominality and length are the decisive predictors on the choice of dative variant with *give*. Subtle but significant variation in the strength of these two factors emerge when the authors perform a by-variety comparison (Schilk et al. 2013: 22).

Cross-varietal differences are not only observable on the probabilistic level but also with respect to distributional patterns and lexical preferences. Regarding distributional patterns, comparative work shows that the prepositional dative is more frequently used in South Asian varieties of English and Canadian English compared to British English (see Olavarría de Ersson & Shaw 2003; Mukherjee & Hoffmann 2006; De Cuypere & Verbeke 2013; Schilk et al. 2013; Bernaisch et al. 2014; Tagliamonte 2014). Regarding lexical preferences and complementation patterns, studies conducted by Mukherjee and colleagues highlight the diversity in verbal preferences among different Southeast Asian varieties and between British English and non-native varieties. For instance, Mukherjee & Gries (2009) assess the strength of the association between individual verbs and monotransitive, ditransitive and intransitive constructions using collostructional analysis. They conclude that British and Hong Kong English, and Indian and Singapore English behave similarly in collostructional preferences. However, the verbs that prefer ditransitive constructions vary across all four varieties (see also Schilk et al. 2013). Mukherjee & Gries (2009) attribute this difference in preference to deviations in the underlying basic ditransitive pattern (see also Mukherjee & Hoffmann 2006: 158). These deviations occur because of differences in

the range of verbs that are used in the ditransitive dative (see Mukherjee & Hoffmann 2006). Mukherjee & Hoffmann (2006: 161) list 20 innovative low-frequency dative verbs in Indian English (e.g. *advise, brief, gift, impart, put, remind, rob* and *inform*, among others) that are not found in the ditransitive dative variant in British English. As Colleman & De Clerck (2011) show, some of these verbs, such as *inform*, originate in the ditransitive dative in eighteenth-century British English and were preserved over time by speakers of Indian English while being lost in Standard British English.

Various explanations have been proposed to account for differences in verb–complementation patterns, in distributional preferences and in probabilistic constraints. Regarding verb-complementation patterns and distributional preferences, differences have been interpreted in terms of cultural factors and substrate effects. For instance, Olavarría de Ersson & Shaw (2003) suggest that the different profiling of recipient/goal and theme with certain verbs, for instance *provide*, might be the result of cultural dissimilarities between Britain and India. Also asserting a cultural explanation, Mukherjee & Hoffmann (2006) explain the preference for the prepositional dative in Indian English with the high frequency of *give* in light verb constructions (see also Hoffmann et al. 2011). De Cuypere & Verbeke (2013), on the other hand, argue that the high number of prepositional dative variants in Indian English is most probably the result of a transfer from Hindi where a similar dative structure requires an explicit dative case marker (*ko*) and where compound verb constructions with *give* are quite frequent (De Cuypere & Verbeke 2013: 181). Differences in the strength of probabilistic constraints across regionally distinct varieties (for instance, between American and New Zealand English, and between American and Australian English) have been attributed to the usage- and experience-based nature of language: As successive generations of speakers are exposed to subtly different linguistic input, gradient differences emerge between their grammars (Bresnan & Hay 2008: 255-256; Bresnan & Ford 2010: 204).

Even though the malleability of probabilistic constraints impacting the choice of dative variant across regionally distinct speech communities has thus been repeatedly empirically (at)tested, the subtlety of these differences has led the majority of studies to argue for homogeneity in probabilistic grammars (see, e.g. Schilk et al. 2013; Bernaisch et al. 2014). Comparing the strength of constraints on dative choice in British and Canadian English, Tagliamonte (2014: 313) finds hardly any differences between these two varieties. The same degree of homogeneity is observed by Bernaisch et al. (2014) across South Asian varieties of English (Bangladeshi English, Indian English,

Maldivian English, Nepalese English, Pakistani English and Sri Lankan English). The results of their study indicate "that many of the predictors found to be relevant in British English are at play in the South Asian varieties too" (Bernaisch et al. 2014: 7). Bernaisch et al. (2014) even go so far as to conclude that the factors determining the choice of dative construction "seem to form part of the 'common core' [...] of English lexicogrammar" (Bernaisch et al. 2014: 28; see also Quirk et al. 1985: 16). This prevailing homogeneity has also been observed on the local level. For instance, Kendall et al. (2011) detect no significant differences in their two models sampling data from African American Vernacular English and General American English and conclude that the dative alternation is not socially variable.

Despite the considerable amount of work on the dative alternation, the full extent of constraints' cross-lectal plasticity is still not well understood. For one, earlier studies often focused on the prototypical verb *give* and only rarely did they consider the full range of ditransitive verbs (for the latter see Bresnan et al. 2007a; De Cuypere & Verbeke 2013; Wolk et al. 2013). Second, the scope of varieties studied was limited to either one or two varieties (e.g. Mukherjee & Hoffmann 2006; Bresnan & Hay 2008; Bresnan & Ford 2010) or to a regionally close group (e.g. Schilk et al. 2013; Bernaisch et al. 2014). The present study's large-scale comparative perspective will thus fill the gap in our current understanding of the cross-lectal malleability of probabilistic constraints that fuel variation in the English dative alternation.

2.7 Chapter summary

This chapter has introduced earlier studies on the English dative alternation from a diachronic as well as synchronic perspective. Work on the history of the ditransitive and prepositional dative shows that the prepositional dative and ditransitive dative were lexically quite restricted in Old English but encroached on each other's turf over time. Especially the prepositional variant became increasingly used with verbs that originally only occurred in the ditransitive dative thus giving the prepositional pattern a boost.

The overview of the synchronic work has presented research on the dative alternation from the perspective of generativism, Construction Grammar, language acquisition research, Variationist (Socio-)linguistics, Probabilistic Grammar and finally World Englishes. Pervasive in this body of research is the separation between those scholars who assume the two variants to be semantically equivalent (the single-

meaning approach) and those who argue for distinctiveness (the multiple-meaning approach). With the upsurge of usage-based approaches in linguistics and the availability of objectively searchable production data, the research focus largely shifted from attempts to prove semantic equivalence to studies first and foremost interested in the underlying constraints that shape variation between variants that are interchangeable under the same truth conditions. The bulk of these studies show that the factors constraining dative choice are multifaceted and simultaneously affect the choice of variant. Cross-regional comparisons across World Englishes further illustrate that these factors are overall fairly stable globally but are also prone to some degree of probabilistic indigenisation.

Since all of these probabilistic approaches to syntactic variation in World Englishes have been restricted in some ways – be it that they only focused on the verb *give*, or on one or two, or a regionally defined set of varieties – the present study constitutes one of the necessary next steps that offers a more comprehensive account of the cross-lectal plasticity of probabilistic grammars.

3

Englishes around the world

3.1 Introduction

In order to properly contextualise regional variation in probabilistic grammars across varieties of English, those varieties' socio-historical backgrounds first need some fleshing out. Nine varieties will take centre stage in the present analysis, namely Canadian English, British English, Hong Kong English, Indian English, Irish English, Jamaican English, New Zealand English, Philippine English and Singapore English. The selection of these nine varieties is largely due to the availability of the corpora from which the data were drawn (see Chapter 4).

The current chapter is structured as follows: Section 3.2 introduces models of World Englishes that categorise varieties of English on regional, historical or evolutionary grounds. Sections 3.3 to 3.11 sketch the socio-historical background of each of the nine varieties. In addition, and if applicable, each socio-historical description is followed by a brief summary of earlier work that investigates the dative alternation in that specific variety. The varieties are presented in alphabetical order.

3.2 Comparing Englishes world-wide

Two research foci can be identified in the study of World Englishes: The first focus pays attention to the distinctive socio-cultural and linguistic dimension that describes one variety's phonological, lexical and/or morphosyntactic make-up (e.g. Kortmann et al. 2004). The second focus adopts a bird's eye perspective and addresses similarities and

differences between varieties in order to classify them into broader types. Research with this second focus often relies on or devises models of World Englishes and has increasingly gained momentum in the past few years (see Kachru 1985; McArthur 1998; Schneider 2007; Mesthrie & Bhatt 2008; Melchers & Shaw 2011; Mair 2013; Buschfeld et al. 2014; Schneider 2014). Studies concerned with the categorisation of World Englishes into different types most basically distinguish between native mother-tongue (L1) or ENL varieties (e.g. New Zealand English), non-native indigenised second-language (L2) or ESL varieties (e.g. Hong Kong English) and English-based pidgins and creoles (e.g. Jamaican Creole) (see Kortmann & Lunkenheimer 2012). Irish English, which is a so-called language-shift variety is often subsumed under L1 varieties due to its long existence under British English influence. Against this backdrop, linguists aim to establish those linguistic features that are particularly diagnostic of a certain variety type, paying attention to phonological, lexical and grammatical variation within and across New Englishes. In that regard, earlier work has mainly described and compared varieties of English in terms of the variable absence or presence of certain linguistic features (e.g. double negation), or in terms of these features' usage frequencies (see, e.g. Kortmann et al. 2004; Kortmann & Lunkenheimer 2012). By comparing varieties' socio-linguistic history, scholars have then tried to identify the common linguistic and socio-historical denominators and classify varieties of English into broader types along different dimensions besides the classical distinction of native vs. non-native vs. pidgin and creoles (see Siemund 2013: 5). However, none of these broader categorisation studies have zoomed in on the underlying constraints that govern linguistic choice making in varieties of English. The question then, whether language users' grammatical knowledge differs across varieties, has so far remained unaddressed.

While it is not the aim of the present study to classify varieties of English based on the underlying constraints that govern syntactic variation in the dative alternation, it is nevertheless necessary to sketch the socio-historical settings of these varieties in more detail in order to properly approach the regional variation observed among speakers' probabilistic grammars.

3.2.1 ENL-ESL-EFL

Probably the most well-known (and still widely employed) distinction between variety types was proposed by McArthur (1998). His model distinguishes between ENL

(English as a Native Language), ESL (English as a Second Language) and EFL (English as a Foreign Language) countries along a regional dimension. The distinctions are quite static: The model views each English as a decontextualised linguistic system spoken in a territorial limited speech community (Mair 2013). ENL territories are those regions where English is first and often the only language of its speakers, such as the British Isles, New Zealand, Australia, Canada and so on.

ESL comprises those regions where speakers acquired English in a process of second language acquisition and where individuals' second language competence became socially aggregated to form a phonologically, lexically and grammatically distinct (one might call it nativised) variety of English. English in those countries has become institutionalised in the administrative and educational context (see Mesthrie to appear) and performs a strong intranational function (Schneider 2007: 12). In contrast to ESL countries, EFL refers to those regions where no British or American presence ever existed and where English does not have an official function but where it still holds a large presence in international communication, tertiary education and sometimes the media (Schneider 2007: 12).

Even though the distinction between ENL, ESL and EFL is useful and still widely employed in current research, the model fails to take into account questions of discourse or language ideologies or the multilingual reality of most nations where ENL and ESL speakers live side by side.

3.2.2 The Three Circle Model

Kachru's *Three Circle Model* (1985) mirrors the ENL-ESL-EFL distinction and classifies varieties according to their historical expansion and educational setting (Siemund 2013: 9). Kachru proposes three concentric circles: The *Inner Circle* comprises traditional ENL varieties, that is, the English spoken in England, in the originally Celtic-speaking lands and in the US, as well as the so-called settler Englishes of Australia, New Zealand and South Africa. The *Outer Circle* includes varieties that were formed during the earlier phases of colonialism in non-native settings, where English has become an important institutionalised language and plays a prominent second-language role. These varieties are the non-native Englishes of South and Southeast Asia and of Africa. The *Expanding Circle* varieties include the Englishes spoken in China or Russia where English is recognised as an important international language and taught in school but where the countries lack a history of colonialisation

and English is not used in institutionalised contexts.

Problematic of the ENL-ESL-EFL distinction and to some extent also of the Circle Model are the value judgements that come with it. Recently, scholars have pointed out that both models seem to posit the inner circle varieties, foremost British and American English, as the norm-giving centres which establish forms of correctness, while postcolonial varieties of English are peripheral, deviating from these norms and consequently evaluated negatively. While this might be true for the ENL-EFL-ESL model, Kachru clearly states that a non-native variety might develop a norm-giving status itself (Kachru 1982: 56-57), thus partly allowing for a variety to change its category (see also Buschfeld & Kautzsch 2017).

Other studies have also criticised that both the Circle Model and McArthur's model remain fuzzy with respect to the classification of more problematic varieties. Cases such as multilingual South Africa or Malaysia do not clearly fit any of the categories (Schneider 2007: 14). These varieties, as well as most other emerging varieties of English, are largely influenced by language contact between colonisers and the indigenous population and by L2 acquisition processes (Siemund 2013: 9). While both the ENL-ESL-EFL and the Circle Model provide broad distinctions between varieties that are useful for very general categorisations – especially the distinction between native and non-native varieties – they fail to take the dynamics of a variety's evolutionary process into account. What is more, both models were first proposed in the late 80s and late 90s, allowing us – more than 20 years later – to challenge the suitability of their categorisation regarding the varieties' contemporary situation. A more dynamic categorisation was subsequently proposed in Schneider (2003) and Schneider (2007).

3.2.3 The Dynamic Model

Inspired by Mufwene's (2001) language evolution theories, Schneider (2003, 2007) provided a unifying model of World Englishes that accounts for the diachronic evolution of new varieties of English while also allowing a more dynamic categorisation of varieties, the so-called *Dynamic Model*. Schneider's model integrates the ecological settings proposed by Mufwene (see Schneider 2007: 22-24) and applies them to the evolutionary dynamics of emerging new varieties of English. In contrast to previous models, the Dynamic Model takes into account theories of language contact, second language acquisition, language evolution and the constant shift and accommodation

in speakers' social and linguistic identity construction, thus factoring in both coloniser as well as colonised populations (Schneider 2007: 21, 31). The model deviates from Mufwene's approach to some extent, for instance with regard to the definition of koinéisation – a term that describes the linguistic process of dialect contact in the early stages of colonial settlement which controversially results either in the levelling of linguistic forms (Schneider 2007: 35) or the "restructuring of a language into a new dialect" without simplifications (Mufwene 2001: 5-6).[1] Essentially, Schneider's model proposes five cyclical stages or phases (see 1.–5. below) through which each emerging variety passes. These five stages are grounded in the socio-historical and -linguistic settings of the colony: When a new dialect or variety emerges, speakers (settlers and the indigenous population) align themselves with other speakers, redefine their social identity through linguistic expression and accommodate their speech behaviour accordingly for communicative purposes (Schneider 2007: 21).

1. *Phase 1 – Foundation*: During the foundation stage, a small number of (mostly British) settlers/colonisers bring their variety/dialect to a new territory. Linguistic contact between new arrivals and indigenous populations remains fairly restricted to utilitarian purposes. Contact between settlers of different dialect origins and between settlers and indigenous populations leads to koinéisation, the borrowing of lexical items describing geographic situations and sometimes incipient pidginisation.

2. *Phase 2 – Exonormative stabilisation*: Increased contact between the indigenous groups and settlers leads to a change in the settlers' identity and bilingual competence in the native population. People of mixed descent play an important role in this scenario: They often act as intermediaries between the local and settler communities. While (British) English remains the reference standard for the written language, lexical items are borrowed from the indigenous language that designate elements of flora and fauna, the culture, customs or other objects that are distinctive of the native community.

3. *Phase 3 – Nativisation*: The third stage is the central phase "of both cultural and linguistic transformation" (Schneider 2007: 40). Growing linguistic and political independence from the motherland and intensified contact between the former settlers and the native population lead to mutual accommodation and

[1] Note, that Mufwene's approach will not be discussed in detail here but see (Mufwene 2001).

an increase in linguistic markers to signal this new identity. At the same time, differences in cultural backgrounds, ethnicity and language often persist, with the pressure of accommodation primarily resting on the indigenous population. Social class will play a major role at this stage since increased contact with the native community will occur at the lower end of the social stratum and in informal communication settings. The tendency to accept localised forms increases gradually and the difference between the former settlers and the native population is often reduced to a social class differentiation. It is at this stage that *structural nativisation* can be observed, that is, the emergence of locally distinctive linguistic forms and structures.

4. *Phase 4 – Endonormative stabilisation*: Even though the transition between phase 3 and 4 can be smooth, phase 4 is often triggered by what Schneider (2007: 49) calls 'Event X' – a (socio-)political event that triggers mental independence from the former motherland and gives rise to a new nation and consequently a new national linguistic identity. The acceptance of local linguistic forms and norms often results in literary creativity in the new variety and in the production of local usage guides and national dictionaries (as, for instance, in Australia and New Zealand). Differences between the settlers and indigenous groups will be less pronounced although they will persist along ethnic and social class lines, while most speakers of the latter group will have undergone a process of language shift.

5. *Phase 5 – Differentiation*: In the last stage, the mostly homogenous national variety of phase 4 will have evolved into an externally stable variety, making room for internal differentiation regarding social, economic and personal status. New group identities emerge based on the speakers' dialects. At the same time, old ethnic boundaries (can) resurface. According to Schneider (2007: 54), the extent of dialect differences emerging in this last stage is a function of the amount of bilingualism developed in phase 4. Differentiation in this final phase is primarily regional and not social since some social variation will have persisted throughout all five stages.

Schneider (2007: 55) cautions that the model "represents a generalization which abstracts from many complexities and details and which captures and highlights certain aspects of reality which are believed to be essential and insightful". Since its publication in 2003 and 2007, other researchers have adopted Schneider's model and

tested it against new and old data. These new studies illustrate that some varieties have progressed towards later stages since the model was first proposed. Hong Kong English, for example, will arguably move into phase 4 in the near future (Setter et al. 2010: 112-116; see also Schneider 2014: 13). Singapore English seems to have progressed into stage 5 (Wee 2014), while Mukherjee (2007) places Indian English in phase 4. Philippine English, which Schneider argues to have become fossilised in phase 2, is actually displaying characteristics of phase 3 and even phase 4 (Pefianco Martin 2014: 74, 78-81). So, what is the current state in the nine national varieties under scrutiny here?

3.3 British English

English in the British Isles emerged out of a mix of Germanic and Celtic ancestry, Norse invasions and Norman-French occupation (Kortmann & Upton 2004: 28). Even though Britain entered the age of exploration and of colonialism relatively late, it nevertheless soon gained political and economic power in India, Southeast Asia, the Americas, Africa, Australia and the Pacific. The British brought not only their goods (and guns) to those parts of the world but also their language. Eighteenth-century British English thus constitutes the input for all but one variety under investigation in this study. In Ireland, Canada and New Zealand, British English was the settlers' language and quickly dwarfed the use of any native tongues. In India, Singapore, Hong Kong and Jamaica, British English was first introduced in the context of trading posts and, in the case of Jamaica, slavery. For a long time, British English remained largely restricted to elitist usage in the colonial context. Only after the Second World War and India's independence in 1947 did the British Empire allow the teaching of English to the masses. The Philippines, which had become an American colony in 1898, escaped the grasp of the British.

The British Isles have been a prolific field of study for dialectologists interested in grammatical variation (e.g. Trudgill 1984; Milroy & Milroy 1993; Tagliamonte & Smith 2002, 2005; Bresnan et al. 2007b; Szmrecsanyi 2008; Haddican 2010; Tagliamonte & Baayen 2012; Szmrecsanyi 2013; Tagliamonte et al. 2014) and especially for those focusing on the English dative alternation (see Gast 2007; Siewierska & Hollmann 2007; Gerwin 2013, 2014) due to the dialectal (non-standard) patterns found here. Two non-standard dative variants are used by British English dialect speakers in addition to the ditransitive and prepositional patterns introduced in Chapter 2, namely

the alternative ditransitive variant as in (3) and the alternative prepositional variant as in (4). The latter is also often referred to as Heavy-Noun-Phrase Shift: The recipient is shifted to immediate postverbal position because of a long and 'heavy' theme (see Gerwin 2014: 5-6).

(3) alternative ditransitive dative
    a. *John gives the apple Mary*
    b. *John gives it her*

(4) alternative prepositional variant
    a. *John gives to Mary the apple*
    b. *? John gives to her it*

The alternative prepositional variant is attested in English (Biber et al. 1999: 928) but hardly found in dialect data (Gerwin 2014). Because this variant is also often regarded as the result of heavy and long themes, and hence motivated by processing-related rather than regional factors, the focus of dialectologists has primarily been on the alternative ditransitive variant.

Alternative ditransitive variants have declined in usage over the past centuries and are mainly restricted to British English dialects. The alternative ditransitive variant with two pronominal objects is also attested in other regions around the world albeit only rarely (see Gerwin 2013: 446-447 for the use of both *give it me* and *give me it* in contemporary American English; see also Yáñez-Bouza & Denison 2015: 249 fn.). Siewierska & Hollmann (2007) and Gast (2007) were the first to analyse variation in English datives with a special focus on the dialectal, that is, alternative, patterns. Siewierska & Hollmann (2007) analyse variable patterns of dative variants in the Lancashire dialect and argue that the strong presence of the alternative ditransitive variant in this regional dialect, especially when occurring with two pronominal constituents (e.g. *John gives it her*), calls for a broader description of variation in the dative alternation (Siewierska & Hollmann 2007: 96-97). Gast (2007) aims to explain the "paradigmatic mismatch" in some British dialects where the canonical ditransitive dative is favoured with two lexical NPs (e.g. *John gives Mary the apple*) and the alternative ditransitive dative (e.g. *John gives it her*) is preferred with two pronominal constituents (Gast 2007: 32). Gast argues that varieties which display such a mismatch are more conservative since the alternative pronominal pattern originates in Old English while the standard pronominal ditransitive pattern

(e.g. *John gives her it*) constitutes a more novel form. This distinction between conservative and innovative dialects finds support in the fact that those dialects that prefer the standard pattern with two pronouns are spoken in the area corresponding to the historical Danelaw where language-contact might have fuelled an increase of the more innovative pattern (Gast 2007: 52).

The paucity of more comprehensive corpus studies (Gerwin 2014: 68) has recently been met in Gerwin (2013) and Gerwin (2014). In her studies, Gerwin uses FRED (Freiburg English Dialect Corpus; see Hernández 2006) and the spoken section of the BNC (British National Corpus; see Aston & Burnard 1998) to profile the three attested dative variants, that is, the alternative ditransitive dative and the two standard forms, synchronically and diachronically across the British Isles. Her results indicate that the standard ditransitive variant is (by now) predominantly used across all of Britain; a slight preference for the prepositional dative with two full NPs in the South of England seems to diminish gradually (Gerwin 2014: 164-165). Also, the alternative ditransitive dative is prevalent in the North of England. In her work, Gerwin (2014) pays special attention to ditransitives with two pronominal constituents (see also Gerwin 2013). In the case of only pronominal objects, the standard pattern as in *John gives her it* prevails in the north, the alternative ditransitive variant (*John gives it her*) is prominent in the Midlands and the South displays a preference for the prepositional encoding (*John gives it to her*) (Gerwin 2014: 198).

Apart from the interest in dialectal variants, recent work has compared patterns of variation in the standard variants in British English with other varieties of English. Tagliamonte (2014), for instance, shows that the ditransitive dative is increasing in usage frequency in British and Canadian spoken vernacular English albeit to different degrees. Grimm & Bresnan (2009) find a similar increase of the ditransitive dative in journalistic prose, a trend that that is mirrored in American journalistic texts. Using data from the BNC 2014 spoken component, Jenset et al. (2017) report that males prefer the prepositional variant more than females in contemporary British English. The extent to which these patterns of variation in British English reflect similar patterns of variation in former British (and American) colonies remains to be explored.

Finally, two methodological issues need to be addressed regarding the use of British English in a study on variable patterns in World Englishes: First, while we might claim that all postcolonial varieties discussed here experienced the influx of some sort of historical variety of British English (apart from Philippine English), we have to

keep in mind that this input was not homogeneous but composed of numerous social and regional dialects. British settlers and traders did thus not transport a standard British English to these British-ruled regions during the period of colonialisation but rather their own dialects packed with local and social features (see Schneider 2007: 101). And second, due to its input status, British English is often treated as the reference variety in comparative studies to fuel claims about the historical development of a variety. The data used in the present work, however, does not allow us to draw any conclusions about the extent to which a variety has evolved away from its historical input variety but rather offers insights into the extent to which a variety is different from present-day British English. Consequently, the present work will treat all Englishes under scrutiny as varieties in their own right. It remains clearly desirable of future work to attempt a comparison with historical data in order to identify cross-regional differences in the extent to which postcolonial varieties have evolved differently from their input.

3.4 Canadian English

After France had lost the struggle with England over their North American territories in 1763, these territories – part of what today is Canada – officially came into the possession of the British Empire (Boberg 2008: 145). Only a few decades later, immigration from New England to the Canadian territories experienced a boost when British loyalists fled the new American republic after the American Revolution. These early English-speaking immigrants constitute a crucial part of the founding population of Canadian English (see Mufwene 2001: 28-29): Their speech patterns set the norms and the standards to which subsequent immigrant groups had to adjust (Levey 2010: 115). After this massive wave of immigration from the US in the late eighteenth century, Britain actively sent numerous English, Irish and Scottish settlers from the British Isles to Canada to counter an increasing pro-American republicanism in the Canadian colony. Their regional British speech patterns would have diversified the existing dialect mix (Levey 2010: 155). The nineteenth century further saw an influx of immigrants from other European countries (e.g. Jews, Italians, Russians). The demographic development led to a characteristic Canadian English that "shows the effect of a standard Southern British superstratum having been imposed on a North American variety" (Boberg 2008: 148). The connection to Britain remained strong up until the middle of the twentieth century when Canada gradually gained

independence (Schneider 2007: 245). Political nationalism after the Second World War can be directly translated into a growing pride in and consciousness of the unique features of Canadian English. Processes of identity formation, however, still largely fall into a gulf between one's British-loyalist roots and Canadians' American origins (Schneider 2007: 243). Today, Canada strongly identifies itself as a country of immigration and encourages immigrants to retain their roots while adjusting to their Canadian environments. Cultural diversity and linguistic diversification remain important aspects of Canada's linguistic identity and places Canadian English in phase 5 of Schneider's model (Schneider 2007).

On the linguistic level, Canadian English is well-known for its nation-wide homogeneity in all parts of the grammar (see Chambers 2012). This homogeneity is generally attributed to the similar founding population across the whole country as well as to the pressure on newcomers to adapt to the existing norms (Schneider 2007: 247). Consequently, no inner-Canadian differences in the patterns of variation underlying the dative alternation have so far been observed. Patterns of variation in the dative alternation in Canadian English differ, however, from the dative alternation in other varieties, notably British English. Comparing patterns of variation in the dative alternation between Canadian and British English, Tagliamonte (2014) shows that the patterns in CanE have changed over time: Women have retained more prepositional datives than men over time, while the ditransitive dative is overall increasing in frequency. The results of her study further indicate that the prepositional variant is more common in spoken vernacular Canadian English compared to British English (Tagliamonte 2014: 314).

3.5 Hong Kong English

Hong Kong became a British colony in the wake of the First Opium war in 1841-1842 (Mukherjee & Gries 2009: 31). The first exposure to English had already taken place a few decades earlier when contact between the indigenous population and traders from the British East India Company led to the birth of so-called Chinese Pidgin English, which probably played an important role at the very beginning of the new colony (Schneider 2007: 135; see also Setter et al. 2010: 104). Even though English was spreading through missionary schools as a second language – especially after the treaty of 1898 guaranteed Hong Kong economic and political stability as a British crown colony for the next 99 years – the language remained fairly elitist and restricted

to the middle and upper classes (Schneider 2007: 135). Only in the middle of the twentieth century, with the end of the treaty in sight and major social and economical upheavals at hand, education (of English) also spread to the general population where it was increasingly used as a lingua franca for interethnic and international communication (Setter et al. 2010: 105, 106). The years 1980 to 1997 witnessed the growth of a local identity of the Hong Kong people as 'Hong Kong residents of British origin' combining western values with Chinese traditions. Positive attitudes towards English and local identity construction furthered the emergence of a unique variety of English (Setter et al. 2010: 107). At the same time, Hong Kong was characterised by societal bilingualism: The Chinese and English-speaking population remained largely separate from each other with a handful of bilingual Cantonese residents serving as intermediaries.

Today, due to a pervasive bilingual situation and unavoidable structural transfer from Cantonese to English, Hong Kong English has developed distinct linguistic features of its own on the phonological but also lexical, morphosyntactic and discourse levels (Setter et al. 2010: 113). Evidence of structural nativisation places Hong Kong English in phase 3 of Schneider's model even though some remaining traces of exonormative orientation towards British English can still be observed (Setter et al. 2010: 114, 116). While the younger generation increasingly employs code-switching and mixing of Chinese and English as part of their identity construction, attitudes towards the local variety of English are still not persistently positive (Mukherjee & Gries 2009: 32).

3.6 Indian English

British traders of the East India Company and missionaries from England first set foot on the Indian subcontinent in the early 1600s and thus laid the foundations for the linguistic and cultural spread of English in Asia (Schneider 2007: 162). While contact was fairly restricted between English and indigenous languages during the first few decades, the anglicisation of India took off in the middle of the eighteenth century when the British Empire gained political power in that region (Bhatt 2004: 1017). As a result, English was increasingly taught in schools and bilingualism became the norm among the upper classes. English in India gradually acquired a very strong local form as learners in the educational context were not exposed to native speakers of English but to local Indians as teachers (Bhatt 2004: 1018; Schneider 2007: 167). After

independence in 1947, the spread of English rapidly accelerated. English was first supposed to remain an official language until 1965 and then to be replaced by Hindi, a major regional language. The *three-language-formula* that the government issued in 1967, however, called for education in Hindi and English and one other major regional language. The three-language-formula was met with particular resistance by non-Hindi speakers in the South and a discernible indifference towards learning a Dravidian language (the third language in the formula) in the North (Schneider 2007: 166). English remained thus fairly uncontested in its status as the language of the official domain in India. In contrast to many other emerging varieties of English, Indian English does not function as an identity marker but rather as a marker of education, serving as a lingua franca in interethnic communication (Schneider 2007: 167).

Because Indian English is used as the official language in administration, politics, on TV, by the press, in school education and at universities by roughly 35 to 50 million Indians on a daily basis (Mukherjee 2010b: 167), it has been called "the largest institutionalised second-language variety of English" (Hoffmann et al. 2011: 258). This post-colonial variety is characterised by the invention of new lexical, phonological, morphosyntactic and stylistic norms and forms. Mukherjee (2007: 158) argues that "present-day Indian English is characterised both by innovative forces, leading to the emergence of local norms, and by conservative forces, which keep it more or less close to native varieties of English". The consensus prevails that Indian English is marked internally by endonormative stabilisation (phase 4 in Schneider's Dynamic Model; see Mukherjee 2007) and externally by its role as a model of English for neighbouring countries (Hoffmann et al. 2011; Schilk et al. 2013; Gries & Bernaisch 2016). With its spread into the lower stratum of society, mixed forms (e.g. Hinglish) have now become increasingly popular (Schneider 2011: 151).

Verb-complementation patterns in Indian English have received ample attention in recent years (see, for instance, Mukherjee & Hoffmann 2006; Mukherjee 2010a; Schilk et al. 2013; Bernaisch et al. 2014). The bulk of these studies indicate that the prepositional dative is more frequent in Indian English in the same context, that is, with the same verbs, compared to British English (Olavarría de Ersson & Shaw 2003; Mukherjee & Hoffmann 2006; Mukherjee & Gries 2009; Schilk et al. 2012; De Cuypere & Verbeke 2013; Schilk et al. 2013). Furthermore, verb-complementation patterns generally differ in their relative frequencies between Indian and British English, not only with *give* (Mukherjee & Hoffmann 2006; Mukherjee & Gries 2009; Bernaisch et al.

2014) but also with other verbs, such as *provide, supply, entrust, present, pelt, shower, pepper, bombard* or *furnish* (Olavarría de Ersson & Shaw 2003: 155). Differences in verb-complementation patterns are more extensive with ditransitive datives and less so with monotransitive and intransitive constructions (Mukherjee & Gries 2009). Similarly, the ditransitive use of some verbs in Indian English such as *inform, advise, brief, gift, impart, remind* and *rob* constitute not only novel innovations but are also indications of the historical origin of Indian English in eighteenth century British dialects (Mukherjee & Hoffmann 2006: 158; see also Colleman & De Clerck 2011: 17). Finally, Schilk et al. (2012) show that Indian English (and Sri Lankan English) are distinct in the verb-complementation profiles of the verbs *supply, convey* and *submit*. These distinctions thus call into question the unification of Indian and other Englishes of that region as *South Asian Englishes*.

While Indian English is distinct from other varieties of English with regard to distributional proportions of dative variants, the probabilistic constraints that fuel the syntactic variation are largely consistent across South Asian varieties (Bernaisch et al. 2014: 7, 19) and British English (De Cuypere & Verbeke 2013; Bernaisch et al. 2014). Some subtle and statistically significant differences between Indian and other varieties nevertheless emerge regarding the importance of the constraints that influence the choice of variant (Schilk et al. 2013). In contrast to British English (and Pakistani English), the most decisive predictors in Indian English are animacy of the recipient, the semantic class of the verb and pronominality of the recipient (Schilk et al. 2013: 18, 22). At the same time, De Cuypere & Verbeke (2013: 174) observe relative length of the constituents and recipient pronominality to significantly impact dative choice.

In order to account for the diverging frequency patterns of specific verbs and differences in probabilistic constraints, researchers have drawn on cultural and contextual factors as well as substrate effects (see also Section 2.6 on the dative alternation in probabilistic grammars and World Englishes). For instance, Olavarría de Ersson & Shaw (2003) argue that the high frequency of *give* in not-so-prototypical ditransitive constructions in Indian English might be attributed to cultural factors or to the high frequency of light verb constructions with *give* (see also Hoffmann et al. 2011; Elenbaas 2013). Interpreting the difference in probabilistic constraints between British and Indian English observed in their study, De Cuypere & Verbeke (2013) assert that Indian English does not adhere to the tendency of aligning the constituents harmonically (see Bresnan et al. 2007a). Rather, they suggest that the dative alternation in Indian English "may have developed in a manner that differed greatly from the evolution

of the other macro-regional varieties of English [...]" (De Cuypere & Verbeke 2013: 180). De Cuypere & Verbeke (2013: 181) further argue that the high number of prepositional dative variants in Indian English might be due to a compulsory explicit dative case marker (*ko*) in the major substrate language, Hindi.

3.7 Irish English

Ireland experienced its first contact with English in the late twelfth century when Anglo-Norman military leaders from the South West and West Midland of England settled on the island (Hickey 2004: 68). By 1600, these English-speaking military leaders were linguistically completely absorbed by the Irish. In order to reinforce the English presence in Ireland, England started a new campaign of plantations along with the banishment of the native Irish language to the geographical (and also social) margins of the country. Both the newly transplanted British dialects (Scots in the north and West/North Midland varieties in the south) and the absorbed ones from the centuries before left their traces in Irish English (Hickey 2010: 77). The overwhelming dominance of English, both politically, socially and linguistically, ended in a complete language shift from Irish to English by the late nineteenth century. The available historical sources indicate that Irish speakers in the rural areas must have learnt English in a somewhat unguided manner through contact with other English-speaking Irish (Hickey 2010: 80). As in other cases where adult learners shift completely to a different language in an unsupervised fashion, the acquisition process by Irish speakers was primarily guided by their first language, not only on the level of phonology or lexis but also in the syntactic domain.

During the English colonialisation period (1600-1900), a steady wave of Irish emigrants who wanted to escape the economic hardship and religious predicaments in their homeland or who were deported by the English authorities (Hickey 2010: 84) settled in Canada (especially Newfoundland), the US, the Caribbean, Australia and New Zealand, and thus contributed to the dialect mixing in those regions.

3.8 Jamaican English

Jamaica has been part of the English-speaking world since 1655 when the British Empire overthrew the Spanish rule of the island (Beckford Wassink 1999: 58). The first settlement period was marked by the immediate introduction of sugar cultivation,

a tremendous population influx and by a highly variable settler situation (small farms vis-à-vis large plantations). Language contact was ubiquitous – mostly between several British dialects, African languages and older forms of Caribbean English (and possibly even some pidgins). From the late seventeenth century onwards until independence in 1962, Jamaica was a politically and socially stable British colony marked linguistically by the emergence of a basilectal Jamaican Creole (Deuber 2014: 28). The economic situation and the ensuing demand for a growing labour force led to a continuous importation of African slaves who had to adapt quickly to cultural and linguistic norms on the island. English was thereby acquired from fellow slaves. This unguided manner of acquisition had direct implications on the formation of Jamaican Creole (see Lalla & D'Costa 1990; Patrick 2004; Deuber 2014).

Life as a slave labourer was harsh which resulted in a series of major and minor uprisings during the eighteenth and nineteenth century. Although education became available through various steps of emancipation towards the end of the nineteenth century, teaching was slow and only a minority of the population attended school (Christie 2003: 12). The beginning of the twentieth century saw the introduction of a labour movement and the foundation of political parties. The demographic situation in Jamaica was still very much divided: While the formerly British whites had largely developed a strong Jamaican identity from the very beginning of their settlement (see Lalla & D'Costa 1990: 23), the black population could not possibly align with their oppressors in their identity construction of an 'us' and held on to their African culture and heritage (Schneider 2007: 231). Still, some form of loyalty and even local pride arose out of the contact with local fellow slaves and hence fellow Jamaicans (Lalla & D'Costa 1990: 25; see also Schneider 2007: 232).

This division in identity construction is mirrored in the linguistic situation: Black slaves at the lower end of the social stratum spoke mostly only creole and in some cases held on to their native African languages. Blacks of intermediate social status commanded mesolectal speech forms and sometimes also Standard English. White settlers were linguistically still primarily oriented towards the British motherland and commanded both British English and some sort of creole that they had adopted from the slave population.

The end of the Second World War brought democratisation, urbanisation and socioeconomic diversification to Jamaica and led to a growing sense of nationalism (Schneider 2007: 234). A pan-ethnic Jamaican identity, instilled by many to incorporate all classes and traditions, was, however, met with resistance by those who saw

Jamaican Creole as a corrupt form of proper English.

Today, the majority of Jamaicans command a mesolectal form of Jamaican Creole and can avail themselves of most facets of the linguistic continuum – moving extensively (and often consciously) between the basilectal (more creole-like) and the acrolectal (more standard British-like) forms depending on the formality of the situation and the rhetorical effects they want to achieve (see Mair 2002). For a long time, Standard (Jamaican) English was the only language used in official and formal domains, in government business, the schools, mass media and generally in all contexts where written language is required (Schneider 2007: 235), while the use of creolisms in written Jamaican English was clearly signalled as external to the text (for instance, by quotation marks) (Mair 2002: 36). At the other end of the stylistic continuum, Jamaican Creole had (and still has) a perceivable influence on spoken language (Deuber 2014: 27).

The advent of new textual genres in the course of the digital revolution changed this binary stylistic situation, however. The last few decades have witnessed a small but noticeable shift towards a greater acceptance of Jamaican Creole in formal contexts (Deuber 2014: 30-33). Mair (2002) reports that Standard Jamaican English has been moving away from an exonormative orientation towards Standard British English and Received Pronunciation as Jamaicans start taking pride in their dialect as a symbol of their Jamaican identity. Gradually, Jamaican Creole is being used in newspapers, government business and the court system. This growing pervasiveness of Jamaican Creole has an impact on the acrolectal forms on all linguistic levels, even more so on spoken than on written language (Mair 2002: 55).

3.9 New Zealand English

The English language first arrived in modern day New Zealand in 1769 when Captain James Cook claimed the two islands for the British Crown (Bauer & Warren 2008: 39). The first settlers were whalers, traders and missionaries, the majority of whom had come from Australia. Early contact with Maori was extensive resulting in the survival of linguistic forms of the indigenous language up to this day (see, for instance, Holmes 1997). The Treaty of Waitangi in 1840 established British colonial rule in New Zealand and migration became more systematic and widespread thereafter (Bauer & Warren 2008: 39). Generally, three waves of immigration are distinguished: In the first wave numerous organisations brought people from London and the South East England,

from Devon and Cornwall and from Scotland to different planned settlement pockets across the islands. The second wave was triggered by the gold rush in the 1860s which led to a large increase in population size with immigrants mostly arriving from Australia and Ireland. Finally, the third wave (starting in the 1870s) brought settlers from southern England to New Zealand. It is generally believed that by 1890, when New Zealand-born English speakers started to outnumber the new immigrants, native New Zealanders became the principal influence on the formation of New Zealand English (Bauer & Warren 2008: 40).

From a linguistic perspective, nativisation was well under way at the beginning of the twentieth century. It is noteworthy that even though New Zealand English had evolved in a short period of time out of a mix of different dialects, historical events and social situations, it remained surprisingly homogenous (Bauer & Warren 2008: 40; Gordon & Maclagan 2008: 64). Australian English and orientation towards the British motherland (politically as well as linguistically) were both influential factors during the formative years, while the Maori population largely shifted completely to English (Schneider 2007: 130).

Ties to the British homeland were loosened with the Dominion status in 1907 and full independence in 1947, although the bond to Britain remained strong (as shown by New Zealand's participation in various wars on the side of the British). When the United Kingdom joined the European Union in 1973, New Zealand lost an almost exclusive export market and had to re-orient itself toward the Asia-Pacific region (Schneider 2007: 131). Linguistically, this event triggered endonormative reorientation and resulted in a locally rooted identity construction of New Zealanders (Schneider 2007: 131). The new identity construction further led to the advent of national dictionaries, separate grammar books, literary creativity and generally a wave of codification. Recent studies show that present-day New Zealand English can be situated at the threshold to phase 5 in Schneider's model – exhibiting signs but no clear indicators of regional differentiation yet (see, however, Bauer & Bauer 2002).

3.10 Philippine English

Philippine English is unique among the assemblage of the varieties discussed here because its input variety is not British but American English. The United States acquired authority over the Philippines in 1898 as a consequence of the Spanish-American War. After three centuries of Spanish colonialism, the Americans were quick

to promote the spread of the English language as a civilising tool (Pefianco Martin 2010: 247). In 1901, English was made the only official language to be taught in the educational context (Schneider 2007: 140). A few hundred teachers, sent off by the US Senate to teach the indigenous people proper (American) English, successfully advanced the rapid spread of English at the beginning of the twentieth century. In 1937, after the Philippines had received limited sovereignty, the government planned to turn Tagalog into a national language. However, the war years (1939-1945) strengthened English as a symbol of resistance against the Japanese oppressors and it gradually developed distinct lexical innovations and grammatical deviations (Schneider 2007: 141). In the 1970s, the government wished to make both Tagalog (which had by then been renamed 'Filipino'; see Lourdes G. Tayao 2004: 1047) and English mandatory languages to be used and taught in a bilingual educational setting. However, the implementation of Tagalog – a regional language from the southern Philippines – as a national language was met with resistance from the northern Philippines. This left room for English to spread even further into the home and informal context. At the same time, a mixed form of English and Tagalog (Taglish) evolved in the late twentieth century, which "combines the status-related appreciation associated with English with the sociable qualities of Tagalog" (Schneider 2007: 142).

Today, English is the language of formal and public domains used in business, higher education, science and technology, politics, the (print) media and government bureaucracy. It is also regarded as the key to professional advancement and associated with the political elite. Philippine English retains a strong orientation towards American English due to the implementation of the American education system in the early days and as a result of teaching methods (see Pefianco Martin 2010). Proficiency in English was said to have been declining at the beginning of the 21st century. Nevertheless, a gradual shift of Philippine English towards phase 4 (endonormative orientation) in Schneider's model has been observed as more and more Filipinos accept English as their own (Pefianco Martin 2014: 79).

3.11 Singapore English

In 1819, Sir Stamford Raffles founded the Straits Settlement at today's location of Singapore as a trading post for the British East India Company. From the very beginning of Singapore, the settlement was characterised by a kaleidoscopic mix of ethnic groups that spoke different dialects and languages. During that early period,

Baba Malay (Malay spoken by Chinese of mixed Malay and Chinese parentage) and Bazaar Malay (a pidgin variety of Malay) as well as numerous southern dialects of Chinese (e.g. Hokkien, Teochew and Cantonese) contributed to the linguistic context out of which Singapore English eventually evolved (Low 2010: 231). The ethnic composition of Singapore continued to be mixed in later decades: Traders, colonial agents, contract labourers of Chinese and Indian origin as well as other Europeans and Asians migrated to Singapore and participated in its linguistic diversity.

In the early nineteenth century, the British government introduced English-medium schools in order to establish a local English-educated elite in the Straits Settlement (Low 2010: 230). When Malaysia gained independence from British rule in 1957, Malay was declared an official language in Singapore and both English and Malay were given prominence until 1967 as a result of advocated bilingualism (Low 2010: 231). In the educational context, Malay, English, Chinese and Tamil constituted the languages of instruction. During those formative years, the linguistic development of English in Singapore and Malaysia was essentially the same since both regions remained part of the Federation of Malaysia until 1965, when Singapore became an independent nation (Low 2010: 231). After independence, the young nation's economic success and language policy soon gave rise to a modernised and highly industrialised state that incorporates a broadly western orientation with Asian values, resulting in a unique Singaporean identity.

Today, English is used as an interethnic means of communication and is taught in school as a first language to all children irrespective of their ethnicity (Schneider 2007: 156). Singapore English shows visible signs of structural nativisation and has even given rise to a distinctive informal local variety, called Singlish. Despite efforts by the authorities to suppress its spread and use, Singlish – and to some extent also Singapore English – have become identity carriers for (young) Singaporeans to express solidarity and pride with their nation. Literary creativity is flourishing, and attempts at codifying Singapore English have found their expression in the *Times Chambers Dictionary*. There are now even signs of Singapore English moving into Schneider's phase 5 (and beyond, see Wee 2014).

3.12 Chapter summary

The sociolinguistic setting of each variety described above and the classification of those varieties within the broader framework of World Englishes, as outlined in the

models, provide the backdrop against which this study's analysis of regional variation can be placed. Two results from the varieties' descriptions need to be highlighted:

First, the descriptions have illustrated that speakers in a postcolonial context – where English is learnt as a second language – have often both a basilectal and a more standard variety at their disposal (e.g. Singlish vs. Singapore English, Jamaican Creole vs. Jamaican English). The description of the corpora in the next chapter will highlight that the current study analyses patterns of variation in the acrolectal (standard) variety rather than in basilectal speech (see Greenbaum 1996b: 6). Second, the descriptions have emphasised the diversity in the development of all nine varieties. Each variety is fairly unique in the socio-historical setting out of which it has emerged. While not all varieties are clearly posited in one of the phases of Schneider's model but rather move along a continuous evolutionary path, the models of World Englishes still provide useful abstractions and generalisations on which the current study can draw. Table 3.1 offers a very brief summary of each variety's background and its categorisation within Schneider's Dynamic Model and the ENL-ESL-EFL/Circles Model introduced at the beginning of the chapter.

Table 3.1 Socio-historical setting of each variety and categorisation into variety type according to models of World Englishes

Variety	First contact with E.	Languages spoken in formative years	Use of English	Schneider's phase	Variety type
BrE	n.a.	n.a.	everywhere	5	ENL
CanE	1763	French, American/British English	everywhere	5	ENL
HKE	1841	Cantonese, Chinese English Pidgin	official context	3	ESL
IndE	1600s	Hindi, among others	official context	4	ESL
IrE	1600s	Irish (Gaelic)	everywhere	5	ENL
JamE	1655	Jamaican Creole	dependent on speech situation	4	ESL
NZE	1790s	Maori, Australian English	everywhere	5	ENL
PhilE	1898	Tagalog, Spanish, American English	official context	3	ESL
SinE	1819	Hokkien, Cantonese, Malay, Tamil, Punjabi, Creoles	official context, home	4	ESL

4

Methodology

This chapter describes the methodological steps taken for data extraction, annotation and statistical analyses. The corpora from which the data are drawn are introduced in Section 4.1, namely the *International Corpus of English* (ICE) series and the *Corpus of Global Web-based English* (GloWbE). Section 4.2 outlines the extraction and filtering of dative observations following variationist sociolinguistic methodology (Tagliamonte 2006). Section 4.3 reports on the annotation process and describes the factors coded for. The statistical toolkit, that is, the techniques applied in the analyses presented in Chapter 5, is introduced in Section 4.4. Finally, note that a very detailed description of the extraction and annotation process, the dataset itself as well as the R-scripts created for the statistical analyses can be downloaded at www.melanie-roethlisberger.ch/data.

4.1 Corpora

4.1.1 The *International Corpus of English*

The *International Corpus of English* project started out in 1988 with a proposal submitted by Sidney Greenbaum:

> We should now be thinking of extending the scope for computerised comparative studies in three ways: (1) to sample standard varieties from other countries where English is the first language, for example Canada and Australia; (2) to sample national varieties from countries where English is an official additional language, for example India and Nigeria; and (3) to include spoken and

> manuscript English as well as printed English. (Greenbaum 1988; taken from http://www.ucl.ac.uk/english-usage/projects/ice.htm; accessed 16 November 2017)

In his introduction to the volume *Comparing English Worldwide*, Greenbaum called the envisaged compilation of the ICE corpora "an ambitious project". The principal aim of it was to "provide the resources for comparative studies of the English used in countries where it is either a majority first language (for example, Canada and Australia) or an official additional language (for example, India and Nigeria)" (Greenbaum 1996b: 3). Currently, 10 complete ICE corpora sampling acrolectal naturalistic language from regionally distinct varieties of English are available for comparative research, and many more are still on their way (see http://ice-corpora.net/ice, accessed 16 November 2017). In order to qualify for the project, English has to be used in those countries as the major language for communication not only in government administration and educational institution but also among its speakers and in creative writing (Greenbaum 1996b: 4).

Each ICE-component consists of 60% spoken and 40% written data, a total of 500 texts with approximately 2,000 words each, adding up to a 1-million-word corpus for each variety of English. The spoken material contains roughly 300 texts from dialogues (180 texts) and monologues (120 texts) covering a wide range of spoken styles from face-to-face-conversations to scripted and unscripted speeches. The written material contains roughly 200 texts from printed and non-printed sources, covering such text types as letters, student essays, academic writing, creative writing, popular writing and reportage (Nelson 1996; see Table 4.1). The texts are written in educated or standard English: To be included in the corpus, speakers/writers need to have received formal education and completed secondary school or have an appropriate public status (for instance, as politicians or writers) (Greenbaum 1996a: 6).

Due to the restrictions and difficulties that some ICE teams faced in other varieties of English (for instance, in Nigeria and Fiji) and especially because times have changed, certain text types are hard to come by (e.g. letters) and are being supplemented or replaced by similar text types (e.g. emails). Also, in some cases some text types had to be extended quantitatively to compensate for the lack of text in another text type. Each ICE team provides a manual that accompanies the publication of their corpus. The manual describes the decisions made by the ICE compilation team concerning the texts sampled, transcriptions, mark-up and the metadata, and highlights those cases where the process of corpus compilation resulted in deviations from the standard ICE

corpus design.

The current study makes use of 9 of the 10 completed ICE corpora (listed alphabetically). The respective manual is provided in brackets.

- ICE-Canada (ICE-CAN) (Newman & Columbus 2010)
- ICE-Great Britain (ICE-GB) (Aarts et al. 1998)
- ICE-Hong Kong (ICE-HK) (Bolt & Bolton 1996; Bolton & Hung 2006)
- ICE-India (ICE-IND) (Shastri & Leitner 2002)
- ICE-Ireland (ICE-IRE) (Kirk et al. 2007)
- ICE-Jamaica (ICE-JA) (Rosenfelder et al. 2009)
- ICE-New Zealand (ICE-NZ) (Vine et al. 1999)
- ICE-Philippines (ICE-PHI) (Lourdes S. Bautista et al. 2004)
- ICE-Singapore (ICE-SIN) (Nihilani et al. 2002)

ICE-East Africa was excluded from the present study because the corpus design deviates too strongly from the corpus design of the other nine ICE corpora. Note that the abbreviations from the ICE corpora will be used for the coding of VARIETY (see Section 4.3 on the annotation of the corpus metadata). These abbreviations will also be used to designate the varieties in the subsequent figures (i.e. GB instead of BrE).

Table 4.1 Design of the ICE corpora — All ICE corpora share the same corpus structure with only small modifications due to practicalities. Each text contains ~2,000 words. The number of texts is given in brackets for mode, register and genre. The number of texts per text type is given in the column 'No.' (*Source*: http://ice-corpora.net/ice/design.htm).

Mode	Register	Genre	Text type	No.	Label
Spoken (300)	Dialogues (180)	Private (100)	Face-to-face conversations	90	s1a
			Phonecalls	10	
		Public (80)	Classroom lessons	20	s1b
			Broadcast discussions	20	
			Broadcast interviews	10	
			Parliamentary debates	10	
			Legal cross-examinations	10	
			Business transactions	10	
	Monologues (120)	Unscripted (70)	Spontaneous commentaries	20	s2a
			Unscripted speeches	30	
			Demonstrations	10	
			Legal presentations	10	
		Scripted (50)	Broadcast news	20	s2b
			Broadcast talks	20	
			Non-broadcast talks	10	
Written (200)	Non-printed (50)	Student writing (20)	Student essays	10	w1a
			Exam scripts	10	
		Letters (30)	Social letters	15	w1b
			Business letters	15	
	Printed (150)	Academic writing (40)	Humanities	10	w2a
			Social Sciences	10	
			Natural Sciences	10	
			Technology	10	
		Popular writing (40)	Humanities	10	w2b
			Social Sciences	10	
			Natural Sciences	10	
			Technology	10	
		Reportage (20)	Press news reports	20	w2c
		Instructional writing (20)	Administrative writing	10	w2d
			Skills/hobbies	10	
		Persuasive writing (10)	Press editorials	10	w2e
		Creative writing (20)	Novels & short stories	20	w2f

4.1.2 The *Corpus of Global Web-based English*

The *Corpus of Global Web-based English* (GloWbE) is one of the most recent web-derived mega-corpora compiled under the auspices of Mark Davies at Brigham Young University (Davies 2013). The corpus comprises 1.9 billion words from 1.8 million web pages covering 20 different English-speaking countries, including 6 inner circle (American English, Canadian English, British English, New Zealand English, Australian English and Irish English) and 14 outer circle varieties, namely English spoken in India, Sri Lanka, Pakistan, Bangladesh, Singapore, Philippines, Hong Kong, South Africa, Tanzania, Nigeria, Kenya, Malaysia, Jamaica and Ghana. To collect the language material, the corpus compilers first extracted region-specific URLs using Google, sampling both blogs and other (more general) web-based material such as newspapers, magazines, company websites and so on (Davies & Fuchs 2015: 4). Since this process made it impossible to avoid the inclusion of some blogs (roughly 20%) when collecting the URLs from general websites, the complete corpus is divided into about 40% general (more formal) websites and 60% blogs – thus mirroring the spoken/written split in ICE. The compilers then collected the language material from these websites using the URLs and removed any boilerplate, that is, recurring headers, footers and sidebars (Davies & Fuchs 2015: 5). Finally, the compilers tagged the corpus texts with the CLAWS7 tagger. Table 4.2 shows the proportions of text distributed across the nine varieties. Most data comes from British websites, followed by the other inner circle varieties (including India). The outer circle varieties have at least 40 million words of text each (not all varieties are listed here). The proportions somewhat reflect the proportional usage of web-based communication in the sampled countries (Davies & Fuchs 2015: 5).

While the fairly automatic text retrieval clearly aided in the compilation of this mega-corpus, this process also contributes to some of its major drawbacks. Due to the automatic text retrieval, the distinction between blogs and general websites is rather coarse and not always clear-cut. What is more, the social background of language users sampled in GloWbE is unknown, which means that we lack any information on how representative they are of their variety, what their knowledge of English is or their social status. A third disadvantage of GloWbE is the unreliability of the CLAWS7 tagger with informal language (as sampled in GloWbE). Finally, the UK- and US-domains most probably also contain linguistic input of speakers of other varieties than US or British English since, for instance, the English spoken by Indian

Table 4.2 The proportions of words by text type and by country in GloWbE — Only the nine varieties discussed in the present study are listed (*Source:* https://corpus.byu.edu/glowbe; accessed 16 November 2017).

Country	Code	General	Blogs	Total
Canada	CA	90,846,732	43,814,827	134,765,381
Great Britain	GB	255,672,390	131,671,002	387,615,074
Hong Kong	HK	27,906,879	12,508,796	40,450,291
India	IN	68,032,551	28,310,511	96,430,888
Ireland	IE	80,530,794	20,410,027	101,029,231
Jamaica	JM	28,505,416	11,124,273	39,663,666
New Zealand	NZ	58,698,828	22,625,584	81,390,476
Philippines	PH	29,758,446	13,457,087	43,250,093
Singapore	SG	29,229,186	13,711,412	42,974,705
Total all countries		1,300,348,146	583,923,681	1,885,632,973

immigrants would have made it into the British part of GloWbE if the speakers blogged or wrote on a UK-based website (see also Davies & Fuchs 2015 for an overview of the (dis-)advantages of the corpus).

Despite these drawbacks, there seems to be general consensus that GloWbE represents an indispensable asset in the researcher's toolbox (see responses in Davies & Fuchs 2015). Its size allows scholars to carry out a wide range of studies on low-frequency phenomena to explore phraseological, lexical, morphological, semantic and/or syntactic variation across different dialects of English. Hence, small, carefully balanced and annotated corpora such as ICE and mega-corpora such as GloWbE complement each other: While the first might be better suited to look for frequent phenomena (due to its small size) that require a close reading of large parts of the corpus, the latter is better tailored to the investigation of infrequent lexical or morphosyntactic features. Both ICE and GloWbE make use of a shared corpus design which facilitates comparative studies (see Mukherjee & Gries 2009: 34). Note also, that some ICE corpora were compiled towards the end of the twentieth century (or beginning of the 21st). Adding language material from more recent periods (as sampled in GloWbE) might thus offer us insights into the current state of the variety. What is more, adding another register to the sample, namely online blogs and writing, can provide us with more vernacular speech.

4.1.3 Data format

Benedikt Heller tagged each of the nine ICE-corpora with CLAWS7 for part-of-speech (available at http://ucrel.lancs.ac.uk/claws) and regularised the layout of the texts (see also Heller 2018).[1] For GloWbE, I relied on the already existing tagging and only changed the format accordingly. For reasons of feasibility, a randomly selected subset of the full GloWbE corpus was used resulting in ~500,000 words per variety. Attention was restricted to the nine varieties sampled in ICE.

Minor inconsistencies in mark-up of the ICE-corpora were retained and the compilers' mark-up was adhered to whenever available. Due to differences in the input variety of English (American English in the case of Philippine English, British English in all other cases) and because of the heavy influence of the American culture nowadays, spelling of words may vary across and within the nine ICE corpora. As these spelling differences are only relevant for weight issues and only with respect to the measure of constituent length in characters, they are so far disregarded and spelling of the original text is always retained.

4.2 Creating the dataset

Data extraction was done with a perl script which selected all constructions with a verb followed by two noun phrases or pronouns from the CLAWS7 tagged version of the ICE corpora and GloWbE (Section 4.2.1). The extracted tokens were then further filtered to exclude any token that did not constitute a dative token at all (Section 4.2.2) and that was not an interchangeable dative variant (Section 4.2.3). A detailed description of the processes involved in the extraction and filtering of the data are provided in the *Guidelines for the Dative Alternation* (available at www.melanie-roethlisberger.ch/data). The current section presents a condensed version of these guidelines.

4.2.1 Data extraction

All dative tokens were extracted from the corpus using a list of dative verbs adapted from previous literature (Levin 1993; Cueni 2004; Mukherjee 2005; Mukherjee & Hoffmann 2006; Bresnan et al. 2007a; De Cuypere & Verbeke 2013; Wolk et al.

[1] Tagging mistakes with CLAWS7 amount to roughly 5% per variant in ICE, as a preliminary cross-corpus exploratory study shows.

2013). At first, this list contained any verb known to occur in either the ditransitive or prepositional dative variant in Standard English (Levin 1993). Supposedly non-interchangeable verbs in Standard English (e.g. *donate the money to charity* ∼ ∗*donate charity the money*) were included in the list since some such verbs may in fact vary in non-Standard varieties of English. At present, there exist no exhaustive lists of all interchangeable dative verbs in all the varieties studied here.

Next, this list was restricted based on whether a verb was attested in the corpus (ICE and GloWbE). Next, interchangeability of the verbs was tested. A given verb was considered interchangeable if it occurred in both ditransitive and prepositional variants in the ICE corpora or in independent datasets, for instance the full GloWbE corpus (Davies 2013) or in Google indexed by region. If I found at least five instances of the verb in each variant, the verb was considered interchangeable, leaving 86 dative verbs (see 5).

(5) *accord, advise, afford, allocate, allot, allow, answer, appoint, assign, assure, award, bequeath, bid, bring, carry, cause, cede, charge, concede, convey, deal, deliver, demonstrate, deny, devote, drop, e-mail, entrust, explain, extend, feed, flick, forward, get, give, grant, guarantee, hand, impart, inform, issue, keep, lease, leave, lend, loan, lose, mail, name, offer, owe, pass, pay, permit, play, pose, post, prescribe, present, promise, propose, quote, re-allocate, read, recommend, refuse, render, return, sell, send, serve, set, show, sing, slip, submit, suggest, take, teach, tell, throw, toss, vote, wish, write, yield.*

In a second step, I used a lemma list of these verbs in a perl script to extract all dative occurrences (verb lemma followed by two noun phrases or pronouns with an optional *to* between them). Precision and recall of this script was then improved by repeatedly verifying the output of the script against the output of the syntactically annotated version of ICE-GB until it was satisfactory enough to be used to extract dative tokens from the part-of-speech-tagged text files of all nine varieties. The output of that script rendered an enormous number of potential dative tokens which then had to be further restricted in two ways. First, false positives had to be weeded out. False positives are observations that might look like dative constructions on the surface but are, in fact, something else. And second, dative tokens that were not interchangeable, that is, where the alternating variant (ditransitive or prepositional) was not semantically equivalent and grammatical acceptable, were also removed. Both filtering steps are discussed next.

4.2.2 Weeding out false positives

The aim of the first filtering step was to exclude any observation that did not constitute a real dative and which did not contain a recipient and a theme. These so-called false positives follow the surface structure of a dative construction but do not contain a verb, recipient and theme (for instance, *She told me last night.*) (see also Bresnan et al. 2007a; Bresnan & Nikitina 2009; Grimm & Bresnan 2009; Wolk et al. 2013). More information on false positives can be obtained from the *Guidelines for the Dative Alternation*.

4.2.3 Defining the variable context

Next, I defined the envelope of variation following traditional approaches in Variationist Sociolinguistics (e.g. Tagliamonte 2006) and discarded all instances where the other syntactic variant was not grammatically acceptable and semantically similar under the same truth conditions (see the discussion in Section 2.5 on semantic equivalence between the two variants). To exclude non-alternates from the dataset, I adhered to earlier work as closely as possible and based any decision on methodological and semantic grounds (see Bresnan et al. 2007a; Theijssen 2012; Wolk et al. 2013; Tagliamonte 2014)

Non-alternating dative tokens excluded from the dataset include:

(6) instances with intervening prepositional phrases or adverbials
e.g. *[. . .] to* ***send his Finance Secretary with him to New Delhi*** *so that he could release the funds immediately*. <ICE-IND:S2B-003>

(7) instances with particle verbs
e.g. *I must remember to* ***give you your linguistics books back****, Laura.* <ICE-GB:W1B-009>

(8) instances with more than one ditransitive verb
e.g. *[. . .]* ***I give and bequeath all my uh my uh assets to my wife*** *for example okay*. <ICE-CAN:S1B-045>

(9) instances with answers/questions as constituents
e.g. *I mean she she* ***answers yes to everything****.* <ICE-GB:S1B-010>

(10) instances containing quotes or titles from other sources
e.g. *Students occupying a central square chanted "Slobo is Saddam" and were led by a rock band in singing "**Give peace a chance**".* <ICE-GB:W2C-019>

(11) instances that included indigenous words where the exact interpretation of the head noun was unclear
e.g. *Today we are **showing you** <indig>**bharadwaja asana**</indig>.* <ICE-IND:S2A-055>

(12) instances with clausal constituents
e.g. *Mark I was **telling Rachel the deaconess** introduced you to Jean.* <ICE-GB:S1A-028>

(13) instances where the other variant was not semantically equivalent, such as:

- instances involving reflexive pronouns
  e.g. *As a lonely and imaginative person she has **given herself to secret and intense infatuations**.* <ICE-IND:S2B-043>
- instances with verbs that allowed for more syntactic variation than just a binary one
  e.g. *I ask him a question = I ask a question of/to him*
- so-called *to-to*-constructions where the alternating variant would still require a *to*
  e.g. *Thank you again for taking the time to **bring this matter to our attention**.* <ICE-HKE:W1B-028>
- predicative prepositional phrases where the last constituent denotes a change to the first constituent
  e.g. *[. . .] his editor, Maxwell Perkins, has to cut them in half in order to **get them to book size**.* <ICE-CAN:W2B-002>
- instances with spatial goals
  e.g. *My Mom wouldn't let me **bring my presents to school**, [. . .].* <ICE-CAN:W2F-001>
- beneficiary constructions
  e.g. *We **get them uh typed photo copies** uhm uhm just a few of them.* <ICE-IND:S1A-060>

- concealed questions (see Nathan 2006; Aloni & Roelofsen 2012)
  e.g. *and when she **told him the significance of this ritual** he requested the woman to tie the sutra to his wrist also*. <ICE-IND:S2B-037>
- fixed expressions and idioms.
  e.g. *it's **bringing tears to my cheeks***. <ICE-GB:W1B-001>

To determine whether the alternating variant was grammatically acceptable and semantically similar in borderline cases, the alternating variant had to occur at least five times in either Google indexed by region or GloWbE. The latter resource was especially useful in those cases where part-of-speech-tagging facilitated the search. Note that this verification step is different from the one conducted for the verb list: The verification of alternating verbs was necessary to ensure that all possibly alternating verbs were included in the extraction process to begin with; the verification of the variable context was carried out to exclude non-variable context specifically by variety, thereby including instances such as *give birth to sb* in BrE but in no other variety (see the *Guidelines for the Dative Alternation* for more details, especially regarding the classification of fixed expressions and idioms).

In contrast to previous research (see, for instance, Cueni 2004), the final dataset includes imperative verbs, variants with an object followed by a relative clause (generally, all tokens with sentential postmodification are included), coordinated objects and objects with subsequent adjective phrases since such constructions have been shown to be variable. The final number of tokens adds up to N =13,171 across all varieties and corpora (distributions by variety and corpus are given in Table 4.3). The final number of tokens excludes any ditransitive or prepositional variant in which the last constituent is longer than the longest first constituent in the alternating variant, thus excluding tokens where the recipient is longer than 18 words (= the longest recipient in the ditransitive dative) and the theme longer than 23 words (= the longest theme in the prepositional dative).

Detailed information on issues pertaining to the definition of the object boundary, to self-correction, repetitions and intervening hesitation or pragmatic markers (*uh, uhm, you know*) can be found in the *Guidelines for the Dative Alternation*. Note here that in the case of self-corrections, a 'first-come, first taken' approach was applied, that is, I always included the first completed dative token in the dataset. Also, in order to identify the (semantic) head of a constituent – which was necessary for the later annotation procedure – the automatically extracted heads were manually

Table 4.3 Total number of interchangeable dative tokens by variety and corpus — The total number of dative tokens to be analysed amounts to 13,171 datives (do = ditransitive, pd = prepositional).

	ICE			GloWbE			
	do	**pd**	**sum**	**do**	**pd**	**sum**	**Total**
BrE	640	234	874	292	152	444	1,318
CanE	671	250	921	297	139	436	1,357
HKE	841	433	1,274	288	206	494	1,768
IndE	608	471	1,079	304	167	471	1,550
IrE	645	222	867	279	124	403	1,270
JamE	683	260	943	294	138	432	1,375
NZE	736	296	1,032	320	133	453	1,485
PhiE	661	348	1,009	334	162	496	1,505
SinE	772	287	1,059	322	162	484	1,543
Total			9,058			4,113	13,171

verified against the definition of NP heads in Quirk et al. (1985). Thus, if the NP was premodified by a quantifier (e.g. *a litre of water*), the last constituent of the quantifying NP (e.g. *water*) was chosen as head. If the semantic head could not be clearly identified, the first syntactic head was selected. For names of locations or titles of books the full noun phrase was used as head. In most other cases, the last word of the constituent was used as head.

Each dative token was subsequently annotated for a number of conditioning factors that have been shown to impact the choice between the two syntactic variants. In addition to constraints annotated in previous work, I also included novel predictors whose effect on dative choice has not been empirically tested before, such as frequency, lexical density and thematicity.

4.3 Annotation: Predictor variables

This section presents an overview of all conditioning factors that were included in the dative dataset. Some of these predictors pertain to the token itself, that is, the register or context it was taken from. The majority of predictors pertain to the recipient and theme, and a few predictors relate to the verb. Most of these predictors were coded fully automatically using a perl script to do so (for instance, length, definiteness,

complexity or type-token ratio). Only animacy and verb semantics were annotated completely manually. The information provided here is largely based on the annotation guidelines for the project 'Exploring probabilistic grammar(s) in varieties of English around the world' (Grafmiller et al. 2016; see also Heller 2018) and the methodology outlined in Röthlisberger et al. (2017).

4.3.1 Corpus metadata

Information on the corpus metadata was included in the dataset in order to provide unique identifiers at various group levels – for instance, text file, speaker within text and sentence within text – as well as information on corpus, variety, subcorpus, mode and genre (provided by the corpus design). The corpus metadata includes the following predictors in the data frame:

1. **Variety**: English variety of the token ('CAN', 'GB', 'HK', 'IND', 'IRE', 'JA', 'NZ', 'PHI', 'SIN')

2. **TokenID**: Unique identifier for each token in the dataset. Construction tag ('DAT'), followed by the number of the token in the dataset. The numbers are consecutive and start with 1.

3. **FileID**: The file in which the token is found ('ICE-GB:S1A-005' = file S1A-005.txt in the ICE-GB corpus)

4. **TextID**: The number of the text in the file in which the token is found ('ICE-GB:S1A-005:1' = text 1 in file S1A-005.txt of the ICE-GB corpus)

5. **LineID**: The number of the sentence/line in the file in which the token is found. Note that line numbers are not grouped by texts within files but are provided consecutively across texts within the same file. ('ICE-GB:S1A-005:52' = line 52 in file S1A-005.txt of the ICE-GB corpus)

6. **SpeakerID**: The identifier of the speaker within a text. Since written texts do not have speakers, authors of individual written texts are uniformly coded as 'A' ('ICE-GB:S1A-005:1:B' = speaker B in text 1 in file S1A-005.txt of the ICE-GB corpus)

7. **GenreFine**: Fine-grained 12-level distinction in genres corresponding to the file types:

'PrivateDia' (S1A)	'PublicDia' (S1B)	'UnscriptMono' (S2A)
'ScriptedMono' (S2B)	'StudentWrit' (W1A)	'Letters' (W1B)
'AcademicWrit' (W2A)	'PopularWrit' (W2B)	'Reportage' (W2C)
'InstructWrit' (W2D)	'PersuasiveWrit' (W2E)	'CreativeWrit' (W2F)
'blog' (B)	'general' (G)	

8. **GenreCoarse**: Five-level register distinction corresponding to the register distinction provided by ICE and GloWbE. Texts from ICE are coded as 'dialogue', 'monologue', 'printed' or 'non-printed' (see Figure 4.1). Texts from GloWbE are coded as 'online'. There will be no predictor for text types themselves (i.e. 'phonecalls', 'exam scripts', etc.).

9. **Mode**: The spoken/written modality of the token ('spoken' vs. 'written')

10. **Corpus**: The corpus the token was drawn from ('ice' vs.'glowbe')

11. **Subcorpus**: Subcorpus combines the corpus and variety identifier (e.g. 'ice-can', 'glowbe-sin', etc.)

The proportional distribution of dative variants by GENRECOARSE highlight that written texts ('printed', 'online', 'non-printed') show higher proportions of prepositional datives than spoken texts ('monologue', 'dialogue') (see Figure 4.1). Differences among registers are statistically significant at the $p<.001$ level ($X^2(4) = 126.82$).

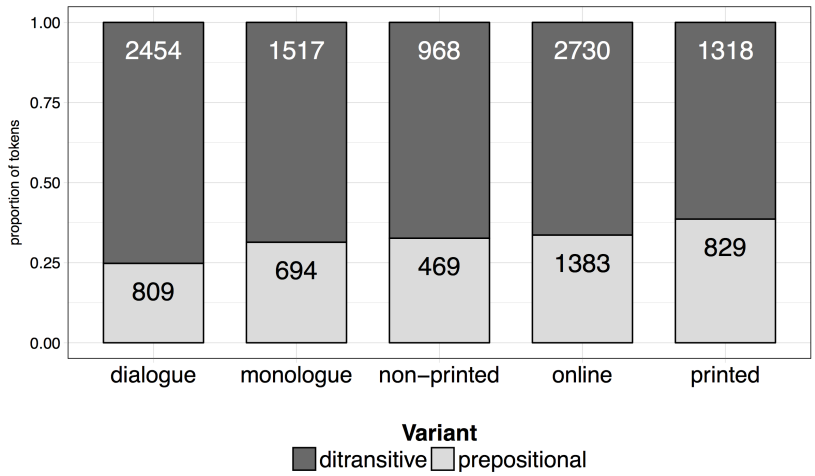


Figure 4.1 Proportion of prepositional and ditransitive dative variants in all five registers (GENRECOARSE) with raw frequencies — Printed texts show the highest proportion of prepositional datives, dialogues the lowest.

The proportional distribution by CORPUS and by VARIETY indicate that the prepositional dative is more frequent in Indian English compared to all other varieties in ICE. In GloWbE, HKE shows the highest proportion of prepositional datives (see Figure 4.2). In both corpora, differences in proportions between varieties are statistically significant (ICE: $p<.001$, $X^2(8) = 132.75$; GloWbE: $p<.01$, $X^2(8) = 21.733$).

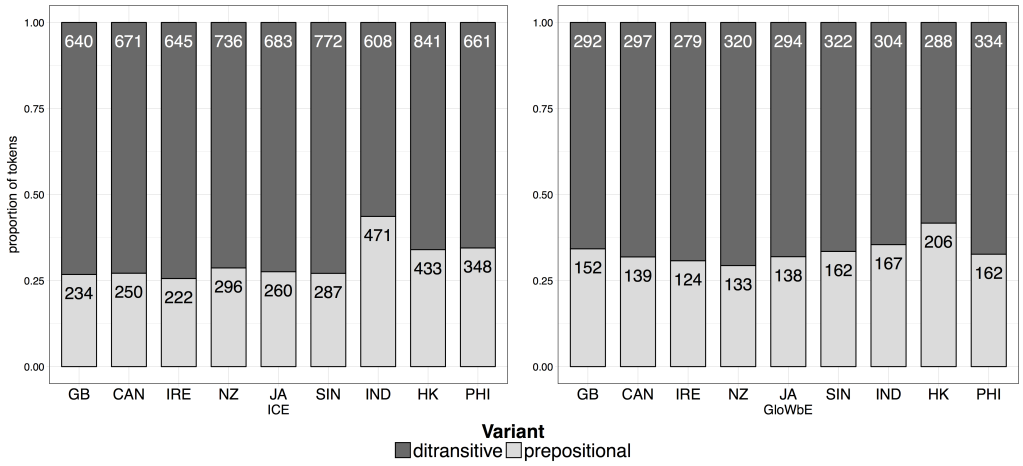


Figure 4.2 Proportion of prepositional and ditransitive dative variants in ICE (left) and GloWbE (right) by VARIETY with raw frequencies — Native varieties appear on the left side (GB, CAN, IRE, NZ), non-native varieties appear on the right side of the graphs.

4.3.2 Animacy

Following Wolk et al. (2013), recipient and theme heads were annotated for their level of animacy using a hierarchical five level distinction. Since animacy seems to have only subtle effects on word order in the dative alternation (see, however, Bresnan & Hay 2008; Bernaisch et al. 2014), I made use of a simplified version of the guidelines in Zaenen et al. (2004) and conflated the five-level distinction to a binary predictor RECANIMACY and THEMEANIMACY with the levels 'animate' (human, animal) and 'inanimate' (all other) (see Table 4.4). Animacy was at first coded automatically using the Manchester Database generated in the project 'Germanic possessive -s'. This database contains, among other predictors, information on the animacy of possessors and possessums sampled from the spoken component of the British National Corpus (see http://www.projects.alc.manchester.ac.uk/germanicpossessive/database/ for more information). The coding provided by this database was then completely manually verified.

Table 4.4 Animacy coding in the dataset — The five-level distinction given here was later merged into a binary predictor RECANIMACY and THEMEANIMACY with the two levels 'animate' (human, animal) and 'inanimate' (all others).

Code	Category	Comments	Examples
'a'	human & animal	only higher animals (not e.g. *fish* or *bugs*); includes spirits, god(s) and other agentive (human-like) supernatural entities	*Shakespeare, engineers, the horse, a sixteen-year-old girl, Mr. Kennedy, God*
'c'	collective	organisations or political states/bodies when seen as having a collective purpose, agenda or will	*the House of Lords, the church, parliament, another country*
		group of animate individuals with potential variable anaphoric reference (*it/they*)	*family, multitudes, the public, a convoy, the majority*
'i'	inanimate	non-temporal, non-locative inanimates: concrete and abstract, all gerunds, participles and infinitives	*the table, oxygen, other topics, drinking*
'l'	locative	places qua places, not groups of inhabitants/members, including *state/empire*; not referable by *they*	*the sea, the playground, China, the earth*
't'	temporal	noun or adverb with time reference	*yesterday, last week, March, 1986, this morning*

Proportional distributions (see Figure 4.3) indicate that animate recipients are more frequent with the ditransitive dative, inanimate ones are more frequent with the prepositional dative. Animate themes, on the other hand, occur more frequently in the prepositional dative (although the number of animate themes is comparatively low) and inanimate themes are more often used with a ditransitive dative. Differ-

ences are statistically significant for both RECANIMACY (p<.001, $X^2(1) = 1401.9$) and THEMEANIMACY (p<.001, $X^2(1) = 33.868$). Effects of animacy align with the expectations given the literature (e.g. Bresnan & Hay 2008; Bernaisch et al. 2014) in that animate constituents are more frequently used in that variant where they occur first and inanimate constituents are frequently used in the variant where they occur last.

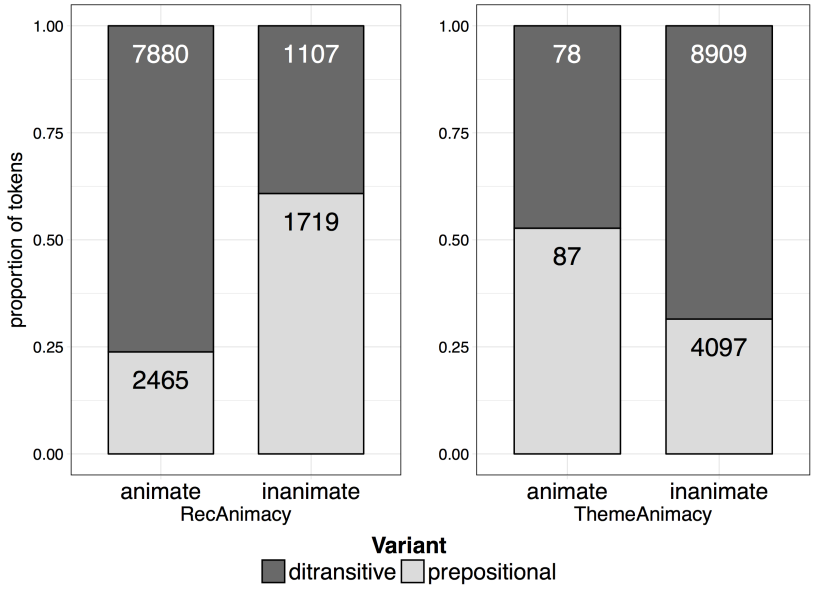


Figure 4.3 Proportion of prepositional and ditransitive dative variants by RECANIMACY (left) and THEMEANIMACY (right) with raw frequencies — Animate constituents are more frequently used in the variant where they occur first and inanimate constituents in the variant where they occur last.

4.3.3 Length

End-weight – often measured in terms of constituent length – is one of the most influential factors when choosing a dative variant (Bresnan et al. 2007a; Gerwin 2014: 48). The term end-weight refers to the general tendency in English to place short constituents before long ones (Behaghel 1909; Hawkins 1994). Two predictors are included in the current study to gauge end-weight effects, namely length and complexity (see next section). The length of each constituent is counted separately in the number of orthographic letters/characters (RECLETTERLTH and THEMELETTERLTH). A few points about these counts are worth noting:

- Spaces are included, while all punctuation is excluded.
- Hyphens are ignored when counting characters.
- Different texts may use acronyms (*NASA*) and initialisms (*U.S.S.R*) in different ways. Since punctuation is excluded when counting letters, differences were minor and ignored.
- Different varieties use different spelling conventions. Some of these are inconsequential for measuring length (*analyse* vs. *analyze*), while others can potentially affect the resulting measurements (*doughnut* vs. *donut*). Variation in spelling across varieties was not corrected.

Note that in addition to length in letters, the dataset also contains a factor that counts the length in the number of words. However, counting the length of constituents in letters provided the more normal distributed data. To reduce multicollinearity in the model and following previous approaches (e.g. Bresnan et al. 2007a), I make use of a log transformed measure called WEIGHTRATIO instead of separate length measurements (see Bresnan & Ford 2010: 174). Taking (14) as an example, where the theme and recipient are 20 and 23 letters long, I calculate the natural log of the weight ratio by ln(recipient length in letters/theme length in letters) = $ln(23/20)$ = 0.140.

(14) *Under the law, LTO should not issue [professional licenses] to [drug addicts or dependents].* <ICE-PHI:W2D-007>

Based on previous literature, we would assume the first constituent in either variant to be shorter than the second, that is, the smaller the weight ratio (<0) the more likely the ditransitive dative becomes, while the larger the weight ratio (>0) the more likely the prepositional dative. This assumption is met by the distribution in the data (see Figure 4.4): The longer the recipient (solid line), the lower the percentage of ditransitive datives, the longer the theme (dashed line), the higher the proportion of ditransitive datives in the data.

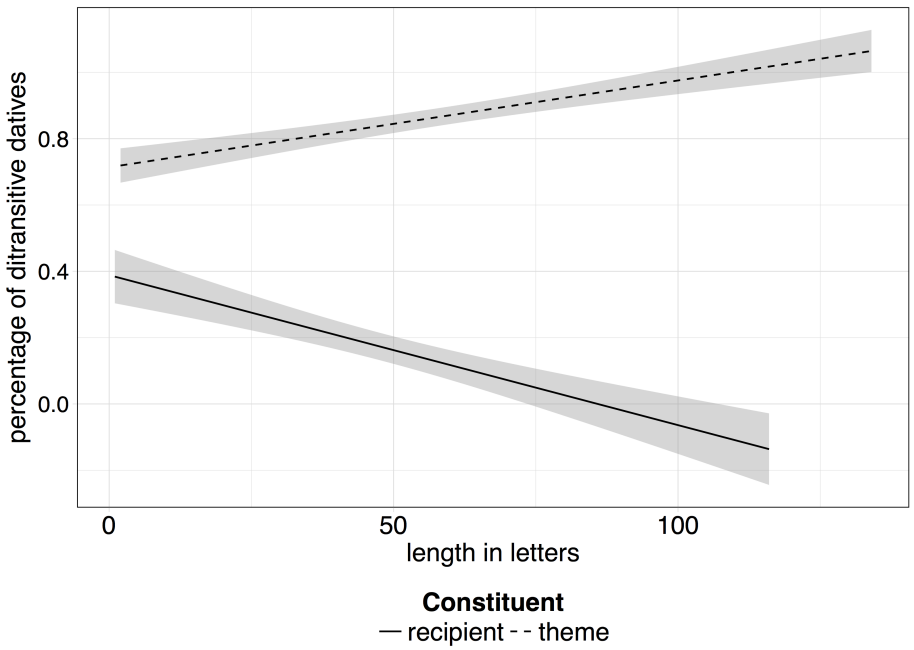


Figure 4.4 Smoothed conditional means of the proportion of ditransitive dative variants by increasing RECLETTERLTH (solid line) and by increasing THEMELETTERLTH (dashed line)

4.3.4 Complexity

The second measure to account for end-weight effects gauges the syntactic complexity of the theme and the recipient. This effect has been shown to constitute an influential determiner of morphosyntactic variation (Berlage 2014) and to be independent of length effects (Wasow & Arnold 2003). For the time being, I coded for a binary distinction between constituent heads with postmodification – coded as 'complex' – and those without postmodification – coded as 'simple' (see example 15). Both the theme (THEMECOMPLEXITY) and the recipient (RECCOMPLEXITY) were coded for complexity.

(15) [...] *they promised [the non-Russian peoples of the vast tsarist empire]*$_{complex}$ *[self-determination]*$_{simple}$. <ICE-SIN:W2E-004>

Given the literature (e.g. MacDonald 2013), we expect simple constituents to precede more complex ones in both the ditransitive and prepositional dative. As the proportional distribution shows (Figure 4.5), simple recipients are indeed more often used in the ditransitive dative than complex recipients and simple themes are

more often used in the prepositional dative than complex themes. The difference in preferences between a simple and a complex constituent is more prominent with the recipient than with the theme. Differences in proportions are statistically significant for RECCOMPLEXITY (p<.001, $X^2(1) = 1706.5$) and THEMECOMPLEXITY (p<.001, $X^2(1) = 1100.5$).

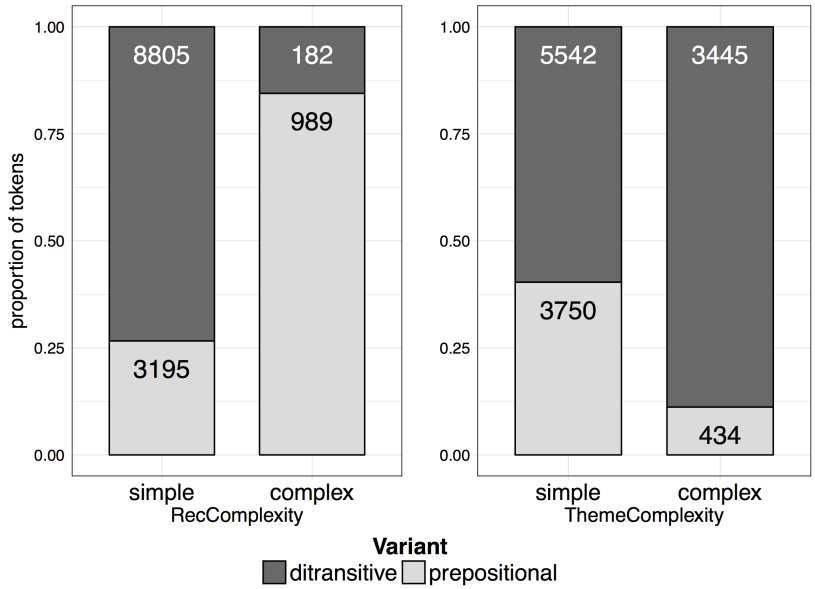


Figure 4.5 Proportion of ditransitive and prepositional datives by RECCOMPLEXITY (left) and THEMECOMPLEXITY (right) with raw frequencies — Simple constituents are more often used in that variant where they occur first compared to complex constituents.

4.3.5 Definiteness

In their experiment with American and Australian participants, Bresnan & Ford (2010) report definiteness and length to be the main factors in their model of the dative alternation (see also Bresnan et al. 2007a). The current study codes themes and recipients for definiteness (THEMEDEFINITENESS, RECDEFINITENESS) following the procedure outlined in Garretson et al. (2004): Any constituent that allowed an existential reading in the context of *There is/are__* (as opposed to a deictic interpretation) was coded as 'indef' (e.g. bare nouns, indefinite pronouns). Additionally, constituents that started with a word marked as indefinite according to Garretson et al. (2004) were coded as 'indef'. These indefinite words include: *a, an, another, any, enough, few, fewer, half, less, little, little or no, lots of, many, so many, more, much, so much, no,*

no more, no such, such, none, one-third, half, one, one or more, ones, plenty of, several, some, twice and *so*. Constituents that contained a proper noun or pronoun as their head or started with a definite article, demonstrative or any word tagged as definite in Garretson et al. (2004) were coded as 'def'. Words tagged as definite in Garretson et al. (2004) include: *the, this, that, those, these, her, his, its, my, our, their, your, all, both, each, either, every, most, neither, last* and *next*.

In accordance with the patterns found in Bresnan & Ford (2010) and others, we expect definite constituents to precede indefinite ones (see also example 16). In other words, a definite recipient should increase the likelihood of a ditransitive dative while a definite theme is expected to increase the likelihood of a prepositional dative. These expectations are borne out by the distribution in the data (Figure 4.6): Definite constituents are more often used in that variant where they occur first compared to indefinite constituents, that is, definite recipients occur more often in the ditransitive dative than indefinite recipients and definite themes occur more often in the prepositional dative than indefinite themes. Differences in proportions are statistically significant for RECDEFINITENESS ($p<.001, X^2(1) = 1018.5$) and THEMEDEFINITENESS ($p<.001, X^2(1) = 110.64$).

(16) *Jim Molyneaux is set to give [the Prime Minister]$_{def}$ [a piece of his mind]$_{indef}$ when the pair meets this week.* <ICE-IRE:W2E-002>

The automatic coding of definiteness raised a problematic issue: As Sand (2004) shows, speakers of L2 varieties tend to use the definite article in contexts where Standard English does not allow it, for instance with generic nouns (e.g. *girls, boys, society, people, men, women*) as in (17) (see Sand 2004: 290).

(17) *The girls tend to fare better in these subjects.* <ICE-SIN:W1A-007>

However, the overuse (or underuse) of the definite article is not a phenomenon restricted to L2 varieties – it has also been observed in English spoken in Scotland, Northern England, South Wales, Ireland and Southwest England (Filppula 1999: 69), as well as in Newfoundland, Singapore, Jamaica, Orkney and Shetland (Siemund 2013: 97). To verify the reliability of the automatic coding procedure for definiteness, I randomly selected 100 tokens (50 tokens with a definite and 50 with an indefinite recipient) from IndE, IrE, JamE and SinE, that is, from those varieties that have been said to show diverging usage patterns of definiteness markers (Filppula 1999; Siemund 2013). The recipient was chosen because Sand (2004) notes that it is animate noun

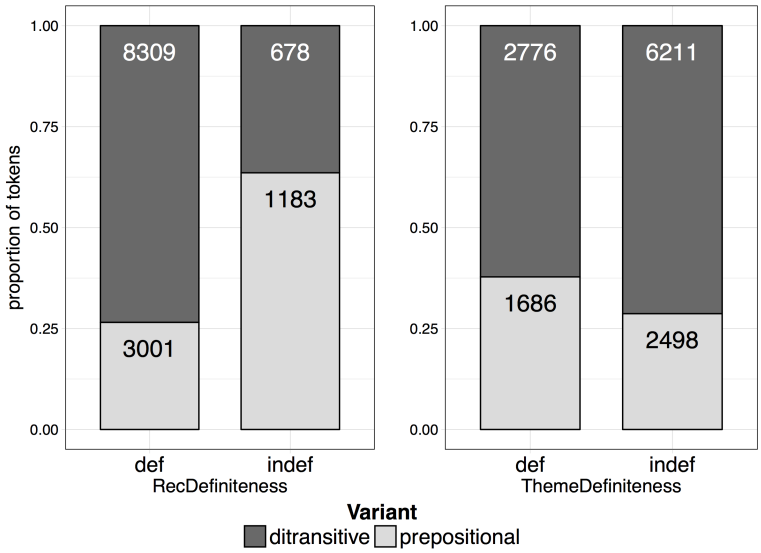


Figure 4.6 Proportion of ditransitive and prepositional datives by RECDEFINITENESS (left) and THEMEDEFINITENESS (right) with raw frequencies — Definite constituents are more often used in that variant where they occur first.

phrases, such as recipients, that tend to be additionally marked with a definite article. After manually verifying the coding for false positives (NPs marked as definite when indefinite) and false negatives (NPs marked as indefinite when definite), I found five miscoded tokens in IndE (mostly cases with *people*), two miscoded tokens in IrE, one miscoded token in JamE and one miscoded token in SinE. While I am thus aware of the complications arising from the automatic coding procedure, the small number of miscoded noun phrases, the unfeasibility of manually verifying over 13,000 tokens and the fact that the non-standard use of definite articles is not unique to L2 varieties all seem to legitimise usage of the automatic coding procedure of definiteness adopted in this study. In any case, any conclusions drawn from the results based on the factors RECDEFINITENESS and THEMEDEFINITENESS will need to be tentative.

4.3.6 NP expression type

To distinguish pronominality (among other things) from the effects of definiteness and givenness (see next section), I added another predictor to the data describing the syntactic category of the relevant constituent heads. To begin with, recipients and themes were coded automatically based on the part-of-speech tag of the constituent head which resulted in a six-level distinction (RECNPEXPRTYPE and THEMENPEX-

PRTYPE) provided in Table 4.5. Due to sparseness of data in some of these levels, levels were merged to create a binary predictor, namely RECPRON and THEMEPRON, which distinguishes between pronominal ('pprn' and 'iprn' now coded as 'pron') and non-pronominal ('nc', 'np', 'dm' and 'ng' now coded as 'non-pron') constituents.

Table 4.5 Six-level coding for NP Expression Type — These six levels were later conflated to two to distinguish between pronominal ('pprn', 'iprn') and non-pronominal ('nc', 'np', 'dm', 'ng') constituents.

Code	Category	Examples
'nc'	common noun	*birds, the market, wisdom, this year*
'np'	proper noun	*President Kennedy, Japan, the United Nations*
'pprn'	personal pronouns, incl. possessives and reflexives	*me, theirs, yourself*
'iprn'	impersonal pronoun incl. *wh* pronouns	*everyone, something, whoever*
'dm'	(bare) demonstrative	*this, that, these, those*
'ng'	gerund (present participle *-ing* forms (rare))	*give your writing a break*

The overall distribution of ditransitive and prepositional datives across pronominal and non-pronominal constituents (see Figure 4.7) reveals that pronominal constituents are more often used in that variant where they occur first compared to the nominal constituents, that is, pronominal recipients are more frequently used in the ditransitive dative and nominal recipients in the prepositional dative ($p<.001$, $X^2(1) = 4268.3$). Pronominal themes are more frequently used in the prepositional dative and nominal themes in the ditransitive dative ($p<.001$, $X^2(1) = 534.28$). This distribution follows the expectations derived from earlier work.

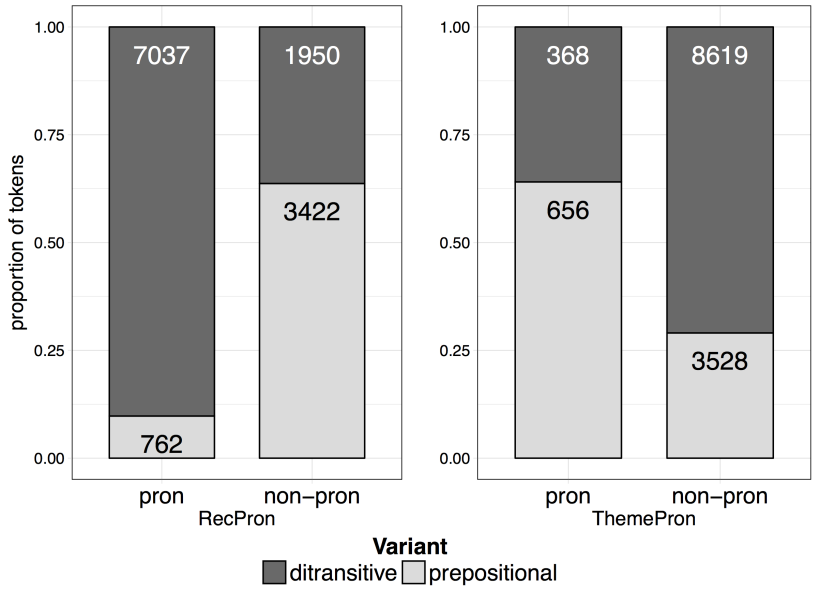


Figure 4.7 Proportion of ditransitive and prepositional datives by RECPRON (left) and THEMEPRON (right) with raw frequencies — Pronominal constituents are more often used in that variant where they occur first compared to nominal constituents.

4.3.7 Information status

Previous research (e.g. Collins 1995; Arnold et al. 2000) has demonstrated that information status exerts an important influence on the ordering of constituents. Following Bresnan & Hay (2008: 249), information status (aka discourse givenness) was coded as a binary predictor ('given' vs. 'new') for both the recipient (RECGIVENNESS) and the theme (THEMEGIVENNESS). If the lemma of the constituent head occurred in the 100 preceding words of discourse or was a personal pronoun, the constituent was coded as 'given'. All other constituents were coded as 'new' (see example 18).

(18) *There is so much that can be got out of story-telling. It is not just to entertain* ***the child*** *but also to feed* ***him*** *with information on* ***his*** *cultural background, to teach* ***him*** *moral values and to enhance family cohesiveness. There are different types of stories and different ways of presenting them. To simplify things, stories could be categorised into family stories and classical stories. Family stories – these stories give [****the child****]$_{given}$ [an idea of himself and the family he belongs to]$_{new}$.* <ICE-SIN:W2D-020>

Findings from earlier work suggest that given constituents precede new ones (Arnold

et al. 2000; Bresnan et al. 2007a). Hence, we would expect given recipients to increase the likelihood of a ditransitive dative and given themes to increase the likelihood of a prepositional dative. This expectation is borne out by the proportional distributions of ditransitive and prepositional datives across given and new constituents (Figure 4.8): Given constituents are more often used in that variant where they occur first compared to new constituents, that is, given recipients occur more often in the ditransitive dative than new recipients and given themes occur more often in the prepositional dative than new themes. Differences in proportion are statistically significant for RECGIVENNESS (p<.001, $X^2(1) = 2940.1$) and THEMEGIVENNESS (p<.001, $X^2(1) = 268.21$).

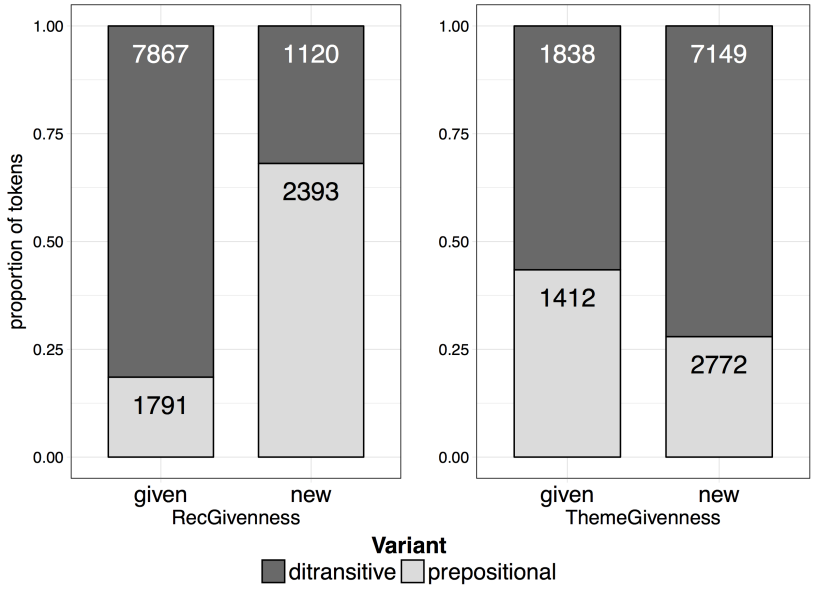


Figure 4.8 Proportion of ditransitive and prepositional datives by RECGIVENNESS (left) and THEMEGIVENNESS (right) with raw frequencies — Given constituents are more often used in that variant where they occur first compared to new constituents.

4.3.8 Verb semantics and verb sense

The literature distinguishes between five broad semantic classes that each dative verb can fall into depending on the context it occurs in (e.g. Bresnan & Hay 2008). Each verb was manually coded according to these five categories (VERBSEMANTICS) and an additional parameter was added (VERBSENSE) that combines the verb lemma and the verb's semantic category for that specific token. For instance, the verb *give* can instantiate three different meanings, namely 'give.a', 'give.t' or 'give.c' which signal

an abstract ('a'), a transfer ('t') or a communicative ('c') meaning of a dative variant with *give*. The five semantic categories are listed in Table 4.6.

Table 4.6 The coding of VERBSEMANTICS in the dataset — Each token's verb semantics was later merged with the verb lemma to create a new predictor, VERBSENSE.

Code	Semantic class	Examples with VERBSENSE in brackets
't'	transfer of possession (of concrete objects)	*They give everybody a piece of paper.* (give.t)
'f'	future transfer (of concrete objects)	*Carl had promised her this car.* (promise.f)
'c'	communication of information	*She told me the whole story.* (tell.c)
'p'	prevention of possession	*They denied him entry to the country.* (deny.p)
'a'	abstract (all other instances)	*You are paying me attention.* (pay.a)

Proportional distributions indicate that dative variants that express 'transfer of concrete objects' and 'future transfer' are more frequently prepositional datives than dative variants that express something abstract, communication or prevention of transfer (Figure 4.9). Differences in proportions are statistically significant at the $p<.01$ level ($X^2(4) = 159.75$).

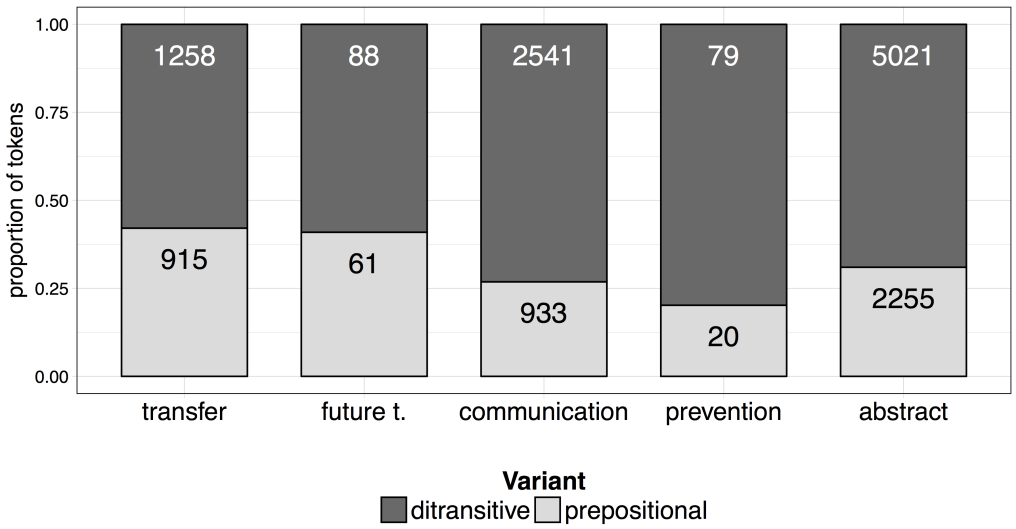


Figure 4.9 Proportion of ditransitive and prepositional datives by VERBSEMANTICS with raw frequencies — Tokens expressing 'transfer of concrete objects' or 'future transfer' are more frequently expressed in prepositional datives than tokens that refer to communication, prevention of transfer or something abstract.

4.3.9 Structural persistence

The effect of persistence – also variably known under the term *structural priming* or *syntactic priming* – refers to speakers' tendency to reuse syntactic constructions that they have heard or uttered previously. Previous work shows that persistence has an effect on the choice of dative variant (Branigan et al. 2000; Gries & Wulff 2005; McDonough 2006) and that this effect might be verb-specific (Gries 2005).

Each dative token was automatically coded for the previous occurrence of dative variant (PRIMETYPE: 'ditransitive', 'prepositional', 'NA') in order to capture effects of persistence/syntactic priming (PERSISTENCE: 'yes', 'no', 'none') (see also Szmrecsanyi 2005). Additionally, after manual verification, the distance to the previous occurrence was used to restrict the effect of persistence to the same text file and the preceding 10 alternating dative tokens. Non-alternating tokens that had been excluded from the analysis were thus ignored for the coding of persistence. For spoken dialogues (ICE corpus), persistence is coded within and across conversation turns, and within and across speakers.

Two predictors thus capture the effect of persistence:

- PRIMETYPE: type of variant used in the previous choice context: 'ditransitive'

for a previous ditransitive variant, 'prepositional' for a previous prepositional variant or 'NA' if the current token is the first one in a text

- PERSISTENCE: indication of whether the priming variant (PRIMETYPE) equals the target variant ('yes') or not ('no'); if no other instance occurred previously in the text, persistence was coded as 'none'.

Proportional distribution points to an effect of structural persistence in the current dataset (Figure 4.10): If the preceding interchangeable dative variant is a ditransitive dative, the ditransitive dative is more frequent than if the preceding variant is a prepositional dative ($p<.01$, $X^2(2)=321.88$).

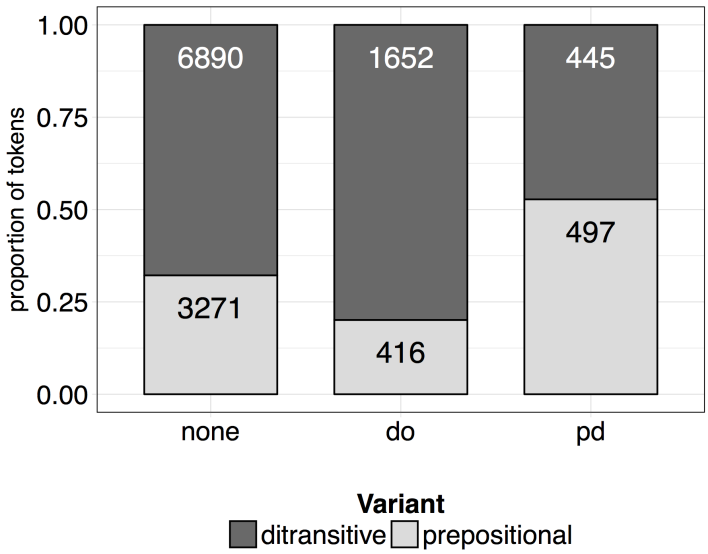


Figure 4.10 Proportion of ditransitive and prepositional datives by PRIMETYPE with raw frequencies — If the preceding interchangeable dative variant is a ditransitive dative (do), the ditransitive dative is more frequent than if the preceding variant is a prepositional dative (pd).

4.3.10 Frequency

Overall frequency of constituent head has been shown to significantly impact phonological and morphosyntactic variation (Gahl & Garnsey 2004; Hilpert 2008) but has never before been included in multifactorial analyses of the English dative alternation. Since we have little information regarding lexical frequency in outer circle varieties

of English, standard lexicons (CMU, CELEX) are not ideal. Consequently, lemma frequencies for each variety were retrieved from the respective complete component of the GloWbE corpus that represents the variety in question. The global frequency of a constituent head (RECHEADFREQ, THEMEHEADFREQ) is normalised as count per million words in the given variety in the GloWbE corpus.

A comparison of means (M) in the current dataset (Figure 4.11) shows that recipients have a statistically significantly higher global frequency in the ditransitive dative ($M = 3064.9$, $SD = 3224.2$) than in the prepositional dative ($M = 742$, $SD = 1849$) as an unpaired t-test shows ($t(12627) = 52.3$, $p<.001$). At the same time, themes have a higher global frequency in the prepositional dative ($M = 1201.5$, $SD = 2631.8$) than in the ditransitive ($M = 422.9$, $SD = 1339.7$). Again, differences in mean theme frequency between the ditransitive and the prepositional variant are statistically significant, as indicated by an unpaired t-test ($t(5218) = 18.1$, $p<.001$) In other words, the more frequent constituents are preferably used in that variant where they occur first.

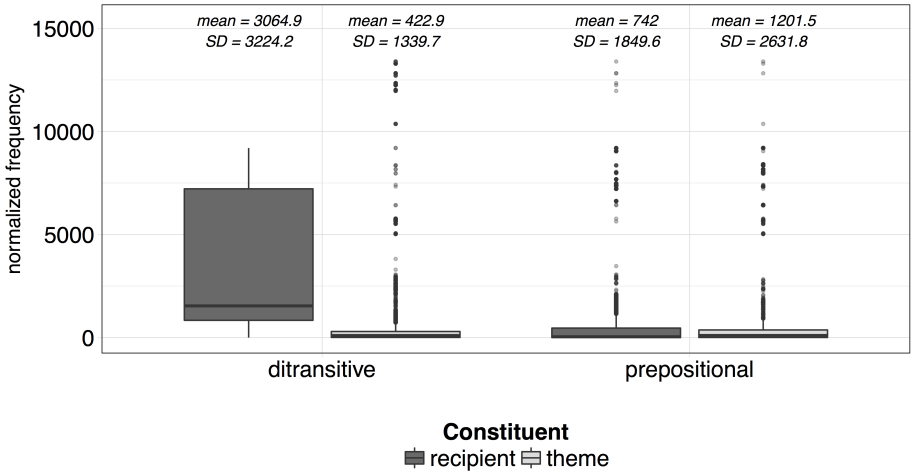


Figure 4.11 Mean (M) and standard deviation (SD) of RECHEADFREQ and THEMEHEADFREQ for each dative variant — In the ditransitive dative (left), recipients are more frequent than themes as indicated by the higher mean. In the prepositional dative (right), themes are more frequent than recipients.

4.3.11 Thematicity

Thematicity reflects the extent to which a constituent forms part of the central topic of a text. Although thematicity has so far not been included in multifactorial studies of the dative alternation, the effect of this predictor has been illustrated in other work on (morpho-)syntactic variation (Osselton 1988). Thematicity is measured here as the normalised text frequency of the head noun in the entire text in which the token occurs, that is, the number of times the constituent head lemma is used in a text divided by the total number of words in the text (i.e. ~2,000 in the case of ICE) (Hinrichs & Szmrecsanyi 2007: 450-451). Both the recipient (RECTHEMATICITY) and theme (THEMETHEMATICITY) were coded for thematicity.

As Figure (4.12) shows, recipients are statistically significantly more thematic in the ditransitive dative (dark boxes) than in the prepositional dative ($t(11962) = 31.5$, $p<.001$). On the other hand, themes are nearly equally thematic in both dative variants (light boxes, $M = 0.003$ and 0.005). The difference in means is statistically significant different as an unpaired t-test shows ($t(5497) = 17.1$, $p<.001$). Consequently, we would expect more thematic recipients to increase the likelihood of a ditransitive dative and more thematic themes to increase the likelihood of a prepositional dative.

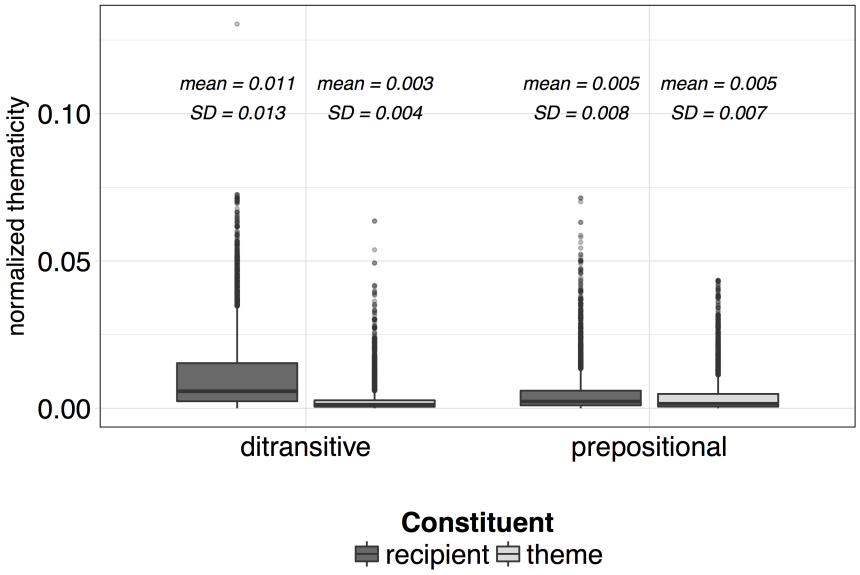


Figure 4.12 Mean (*M*) and standard deviation (*SD*) of RECTHEMATICITY and THEMETHEMATICITY for each dative variant — Differences between variants regarding the means of constituents' thematicity are statistically significant.

4.3.12 Lexical density

The lexical density of the surrounding context of a dative token was gauged using the type-token ratio (TYPETOKENRATIO) of the 50 words preceding and the 50 words following the token. The type-token ratio is defined as the number of unique lemmas divided by the number of word tokens in this 100 word environment surrounding the dative variant in question.

Lexical density has been shown to influence morphosyntactic variation in the genitive alternation (Hinrichs & Szmrecsanyi 2007: 457): Language users seem to meet the need to encode more information economically in a given textual passage with a preference for the *s*-genitive – that variant which "represents a good way of compressing information" (Biber et al. 1999: 302). Following this earlier work, we could speculate that the ditransitive dative – which is the more dense variant in comparison to the more transparently encoded prepositional dative – is more likely in lexically dense contexts. As the data shows, however, increased lexical density requires a more transparent encoding of a dative construction, that is, the more lexically dense the context, the less likely the ditransitive variant (Figure 4.13).

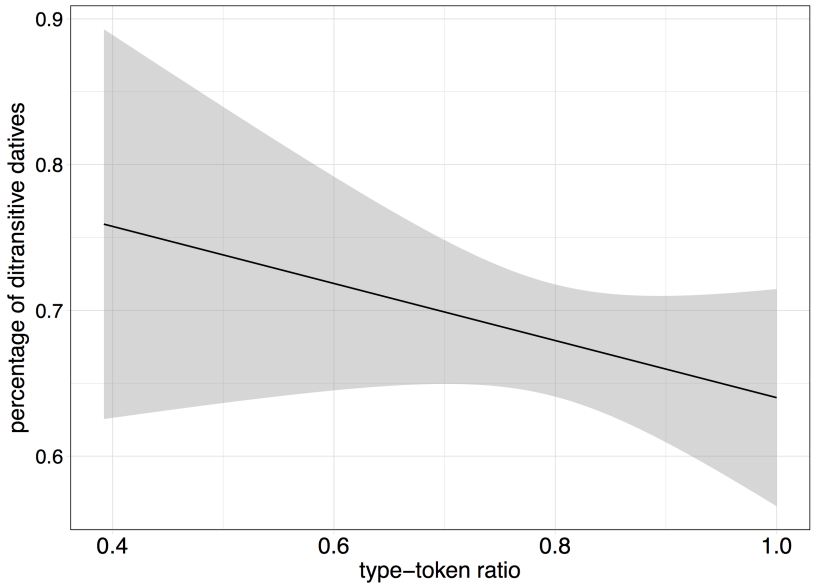


Figure 4.13 Smoothed conditional means of the proportions of ditransitive dative variants by increasing TYPETOKENRATIO — Increased lexical density (x-axis) leads to a decrease in the proportion of ditransitive datives (y-axis).

In sum, the proportional distributions of predictors by the two dative variants highlight

the similar alignment of the predictors' effects. The ditransitive dative is more frequent when the recipient is animate, simple, short, definite, pronominal, given and previously used in discourse, thematic and frequent. At the same time, the prepositional dative is more frequent when the theme is animate, simple, short, definite, pronominal, given and previously used in discourse, thematic and frequent. In other words, and following MacDonald (2013), language users seem to preferably opt for that dative variant where the first constituent is more accessible and thus 'easier' to produce.

4.4 Statistical toolkit

In order to analyse the contribution of the constraints on the choice of dative variant and to gauge their relative importance, the present study employs logistic regression modelling, random forests and multidimensional scaling (MDS).

Logistic regression estimates the simultaneous effect of a set of factors on a binary outcome and gives an indication of the probability of observing one of the variants (Hosmer & Lemeshow 2000; Gelman & Hill 2007). More precisely, I will mainly use mixed-effects modelling which takes not only the combined set of factors into account (as fixed-effect modelling does), but also allows for so-called random effects – by-group idiosyncratic variation that is specific to the dataset (Pinheiro & Bates 2000). Including those idiosyncrasies enables us to better generalise beyond the particular data sample to the population at large. Idiosyncrasies are included in the model as so-called random effects and include but are not restricted to lexical effects (e.g. verb, recipient, theme), text type and the speakers from which the samples originate. Models were fitted with the lme4 package in R (Bates et al. 2015; R Core Team 2016). To evaluate model fit, the prediction accuracy of the model on the same dataset was calculated (given in %) and the *C*-statistic was computed. The *C*-statistic, also known as Somer's *C* index or concordance index *C*, is a measure of how well the model is able to discriminate between the variants of the outcome. Its value corresponds to the proportion of times the model makes a higher prediction for one variant when that variant is also observed in the data. For instance, the proportion of the number of times that the model predicts the ditransitive variant when the ditransitive variant is used, added to the proportion of the number of times that the model predicts the prepositional variant when the prepositional variant is used (Levshina 2015: 259). Hosmer & Lemeshow (2000) propose the following scale to interpret the *C*-statistic and model fit:

	C	$= 0.5$	no discrimination
$0.7 \leq$	C	< 0.8	acceptable discrimination
$0.8 \leq$	C	< 0.9	excellent discrimination
	C	≥ 0.9	outstanding discrimination

Conditional random forests were computed whenever the explanatory importance of the various constraints had to be determined. Conditional random forests seek to predict which of two outcomes (in this case prepositional or ditransitive dative) is more likely given a set of predictors. In contrast to regression models, which make this prediction by specifying how each factor affects the choice on the basis of a mathematical equation, random forests establish the usefulness of a predictor through trial and error by bagging a pre-specified number of conditional inference trees that are computed on randomly selected subsamples (training sets) of the data. Conditional inference trees split the data recursively into smaller and smaller subsets based on those predictors that co-vary most strongly with the outcome. The aim of conditional inference trees is to retain homogeneity in the outcome (e.g. all ditransitive vs. all prepositional datives) in all subsets of the data as much as possible for each binary split. The splitting is repeated until no further splits can reduce the heterogeneity in the data. The prediction accuracy of each tree is then assessed on the not-sampled data or test set and used to evaluate the usefulness of the predictors associated with the splits in the tree. Finally, the importance of each predictor is determined using a conditional permutation scheme on the aggregate estimate of each tree's most likely response outcome (Strobl et al. 2008). Conditional random forests are especially well suited to measure the importance of predictors since the random subsampling and conditional permutation scheme drastically reduce the problem of correlated predictors in a dataset. What is more, conditional random forests overcome common problems of regression models (e.g. data overfitting) in that they can deal with empty cells, with the perfect separation of the response variable in combination with independent factors, and they do not overestimate the influence of numeric predictors or predictors with many levels (see Tagliamonte & Baayen 2012: 158-161 for details). Conditional random forests were fitted using the `cforest()` function in the party package (Hothorn et al. 2006a; Strobl et al. 2007, 2008). In those cases where I fitted conditional inference trees (for instance, to identify possible interaction terms), I made use of the `ctree()` function in the partykit package (Hothorn et al. 2006b; Hothorn & Zeileis 2015).

In addition to these two statistical techniques, I will also make use of techniques traditionally used in dialectometric approaches, namely distance measures and multi-dimensional scaling (MDS) (see, for instance, Szmrecsanyi & Kortmann 2009; Szmrecsanyi 2013). These dialectometric tools are traditionally used to gauge and visualise the distance between dialects or varieties using a frequency-based feature list as input (exemplified in Table 4.7).

Table 4.7 Example of frequency-based feature list — Such lists serve as the input in traditional dialectometric studies to calculate the aggregate distance between varieties.

	Text frequency of feature in Variety A	**Text frequency of feature in Variety B**
feature 1	234	123
feature 2	56	86
...	...	...

As will be shown in Chapter 5, distance measures and MDS can also be applied to quantify the probabilistic distance between varieties using non-frequency based data, that is, coefficient estimates gained from statistical models or constraint rankings obtained from random forests (see also Szmrecsanyi & Hinrichs 2008: 305-306). To calculate the distance between varieties, two different metrics will be used depending on the input data (see Section 5.7 for the application of the metrics to the dative data). Manhattan or City-Block distance calculates the distance or dissimilarity between two objects based on the absolute sum of the objects' vertical and horizontal distances along the gridlines (the dashed line in Figure 4.14). Manhattan distance is thus different from and a special case of Euclidean distance which measures the diagonal distance between two objects (the solid line in Figure 4.14) (Aldenderfer & Blashfield 1984: 25). Euclidean distance is a dissimilarity metric that stems from the Pythagorean theorem. The Pythagorean theorem states that the squared length of the longest side of a triangle (say C) equals the sum of the squared lengths of the other two sides (i.e. $A^2 + B^2 = C^2$). The length of C is thus the root of this sum and equals the linear distance between the endpoints of C, Var A and Var B (see Figure 4.14).

Calculating the distance between objects (varieties in our case) results in a $N \times N$-dimensional distance matrix where N stands for the number of objects to compare. Since more than three dimensions are hard to visualise (conceptually as well as physically), the distance matrix needs to be reduced to a manageable two or three

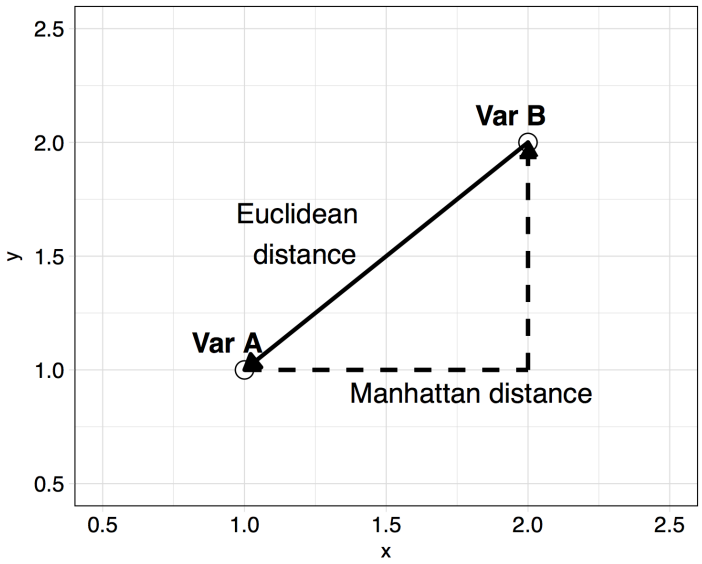


Figure 4.14 Manhattan distance (dashed line) and Euclidean distance (solid line) — Euclidean distance is calculated based on Pythagoras' theorem. Manhattan distance follows the gridlines.

dimensions. Multidimensional scaling is a dimension reduction technique (Kruskal & Wish 1978) and is available in R with the `cmdscale()` function for classical metric MDS (R Core Team 2016) and `isoMDS()` from the MASS package for non-metric MDS (Venables & Ripley 2002). In essence, MDS recreates a new data frame that approximates the original data frame (with the frequency of features) from which then a new distance matrix is calculated but with a reduced number of dimensions. This process is iteratively repeated until the difference between the original distance matrix and the recalculated distance matrix is as low as possible. The aim of MDS is to reduce this difference, also known as *stress*, as far as possible. Since stress increases with a lower number of dimensions, the more dimensions one allows for, the more stress is reduced. There is, however, a cut-off point (an 'elbow' when plotting the reduction in stress) after which no additional dimension reduces stress extensively anymore. Preferably, this cut-off point occurs after two or three dimensions. The stress from a multidimensional scaling model is thus an indication of goodness-of-fit and ranges between 0 and 1, with 0 indicating perfect fit and 1 indicating random noise and no fit at all. To interpret the stress value of MDS, the following rules of thumb can be applied (Levshina 2015: 341):

	stress	> 0.2	=	poor
$0.2 >$	stress	> 0.1	=	fair
$0.1 >$	stress	> 0.05	=	good
$0.05 >$	stress		=	excellent

Earlier work uses frequency-based lists of features, such as the one accompanying the *Handbook of Varieties of English* (Kortmann et al. 2004) to calculate the distance between varieties of English (see also Table 4.7 for an example). The feature list in the *Handbook of Varieties of English*, for instance, provides information on the presence or absence of 76 non-standard features in 46 vernacular varieties of English. Applying MDS to the calculated distance matrix derived from these 76 features renders a two-dimensional plot shown in Figure 4.15. Distances between varieties correspond to their aggregate morphosyntactic dissimilarity. Varieties are grouped based on type (L1 = native varieties, L2 = non-native varieties, PC = pidgins and creoles).

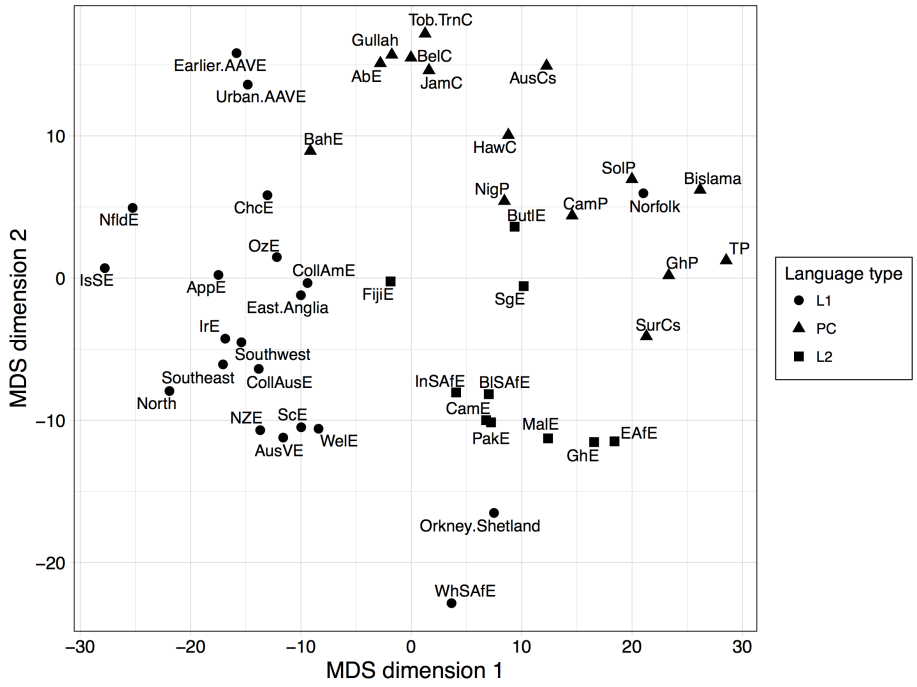


Figure 4.15 MDS map of varieties of English — Distances between varieties correspond to their aggregate morphosyntactic dissimilarity. Varieties are grouped based on type (L1 = native varieties, L2 = non-native varieties, PC = pidgins and creoles) (*Source*: Figure 1 in Szmrecsanyi & Röthlisberger: to appear).

The three techniques introduced here – logistic regression, random forests and MDS –

all form part of the statistical toolkit that will be used for the analyses presented in the next chapter. Where necessary, other techniques will be described in the relevant sections.

5

Regional variation in probabilistic grammars

5.1 Introduction

The current chapter investigates the scope and limits of cross-varietal variation in speakers' probabilistic grammars across regionally distinct varieties of English. The chapter thus aims to address the research questions introduced previously, namely:

- What is the extent to which varieties of English share, or do not share, a probabilistic grammar that is explanatory across different varieties? And what are the limits of cross-varietal variation?
- Are lectal differences random or can they be explained by considering socio-historical factors such as language contact?
- To what extent are factors that are typologically robust cross-lectally variable?
- Which of the individual constraints are tied to stylistic differences or lexical considerations?

To address these questions fully, the statistical methods outlined in Chapter 4 are applied and supplemented with additional techniques where necessary. To determine the relative importance of conditioning factors and to disentangle these factors' simultaneous effect on the choice of dative variant, conditional random forests and

mixed-effects models are fitted to the data. Mixed-effects models and conditional random forests are especially suited to dismantle the multivariate nature of the probabilistic grammar underlying the choice of dative variant since – as previous studies have shown – the choice of dative variant is not influenced by one but multiple, sometimes conflicting and correlating constraints which probabilistically impact syntactic variation.

Sections 5.2 and 5.3 report the results of a conditional random forest and a mixed-effects model fitted on the full dataset. Since the main interest of this study lies in the cross-lectal plasticity of probabilistic constraints on syntactic alternation, regional differences in probabilistic constraints observed in the mixed-effects model are further explored in the subsequent sections. As the model output will show, end-weight, recipient pronominality and corpus vary significantly in their effect size across the nine varieties under scrutiny here. Consequently, Section 5.4 takes a closer look at end-weight effects across varieties of English, Section 5.5 focuses on the effect of recipient pronominality and lexical effects on the choice of dative variant and Section 5.6 investigates the extent to which corpus plays a role by zooming in more closely on the register-specificity of the English dative alternation. Using the effect of corpus as a backdrop to investigate the register-specificity of the dative alternation is justifiable, since corpus distinguishes between ICE and GloWbE and GloWbE samples data from only one specific register (online data). It might therefore be possible that the regional variability of corpus effects is in fact the result of register effects. Finally, the stability of probabilistic grammars and hence limits of cross-varietal variation are quantified in a suggestive attempt based on comparative sociolinguistic methods.

5.2 Establishing relative importance of constraints

5.2.1 Training the forest

To set the stage, a conditional random forest was fitted to the dataset in order to determine the explanatory importance of the various constraints that shape the choice of dative variant. Conditional random forests aggregate over a predefined number of conditional inference trees fitted on randomly selected subsamples of the data and predictors. Inference trees are a classification technique whereby predictors are selected as more important on the basis of how homogeneously they split the data. Conditional random forest then aggregate over these predictors and select predictors

according to their importance in the classification of data points in the conditional inference trees. Due to their random sampling procedure, conditional random forests are quite robust to statistical issues commonly encountered in regression analysis such as data sparseness or predictor non-linearities (see also Tagliamonte & Baayen 2012: 158-161 for details). To calculate the importance of each predictor in the choice of dative variant, the `varimpAUC()` function in the party package was used which calculates the importance of predictors based on the area under the curve (the *C*-statistic) instead of accuracy (Janitza et al. 2013).

The conditional random forest was fitted on the dataset using the `cforest()` function from the party package (number of predictors selected at each split(*mtry*) = 3, number of trees grown (*ntrees*) = 2000) (Hothorn et al. 2006a; Strobl et al. 2007, 2008). The model formula of the conditional random forest includes all predictors introduced in Chapter 4 (shown in 19). Numeric variables, that is, length, head frequencies, thematicity and type-token ratio were scaled by two standard deviations and centred around the mean (following Gelman 2008).

(19) Variant ~ Variety + Corpus + Mode + VerbSemantics + WeightRatio + RecGivenness + ThemeGivenness + RecDefiniteness + ThemeDefiniteness + RecComplexity + ThemeComplexity + RecHeadFreq + ThemeHeadFreq + RecThematicity + ThemeThematicity + PrimeType + RecPron + ThemePron + RecAnimacy + ThemeAnimacy + TypeTokenRatio

Next, I cross-validated the random forest and tuned the hyperparameter (*mtry*) using the `train()` function from the caret package (Kuhn et al. 2016) with repeated cross-validation. This method created 10 splits of the data into training and test sets for a total of three repetitions (Kuhn & Johnson 2016: 71-72). The final and most effective model uses five predictors at each node in the trees of the forest. Another forest was thus run with *mtry* = 5 and 2000 trees to grow. The robustness of this forest was confirmed by fitting the same forest again with a different random seed. If not mentioned differently, all subsequent forests use the same setting (*mtry* = 5, *ntrees* = 2000). The conditional random forest performs well on the data, *C*-statistic is an outstanding 0.95 and predictive accuracy is 89.2% which is significantly better than the baseline of 73.84% of always choosing the more frequent variant (p_{binom}<.001).

The forest generally overpredicts the more frequent variant, as illustrated by the confusion matrix in Table 5.1. For 783 tokens (= 5.9% of all datives), the random forest predicts a ditransitive instead of the observed prepositional dative and in

640 cases (= 4.9% of all datives) the forest predicts a prepositional dative while the observed variant is a ditransitive dative. In other words, the conditional random forest correctly identifies 92.9% of all ditransitive variants and 81.3% of all prepositional variants.

Table 5.1 Confusion matrix of predicted vs. observed variants given the conditional random forest — The random forest overpredicts the more frequent variant.

predicted / observed	ditransitive	prepositional
ditransitive	8347	783
prepositional	640	3401

Conditional random forests fitted separately per variety confirm the over-prediction of the most frequent variant (the ditransitive dative). These forests use the same model formula as (19) apart from the predictor VARIETY and include categorical predictors that were transformed into numbers and centralised. To transform categorical predictors into numbers, each level of the predictor was assigned a number starting from 0 (for instance, for RECANIMACY I assigned '0' to 'animate' and '1' to 'inanimate'). In all varieties except Indian English, the forest models persistently err on the side of the ditransitive dative. Figure 5.1 illustrates this tendency: All varieties, except Indian English, are plotted below the dotted line that represents a perfect correlation of predicted and observed prepositional variants. The reason for this discrepancy between IndE and the other varieties is most probably due to the larger proportion of prepositional datives in IndE.

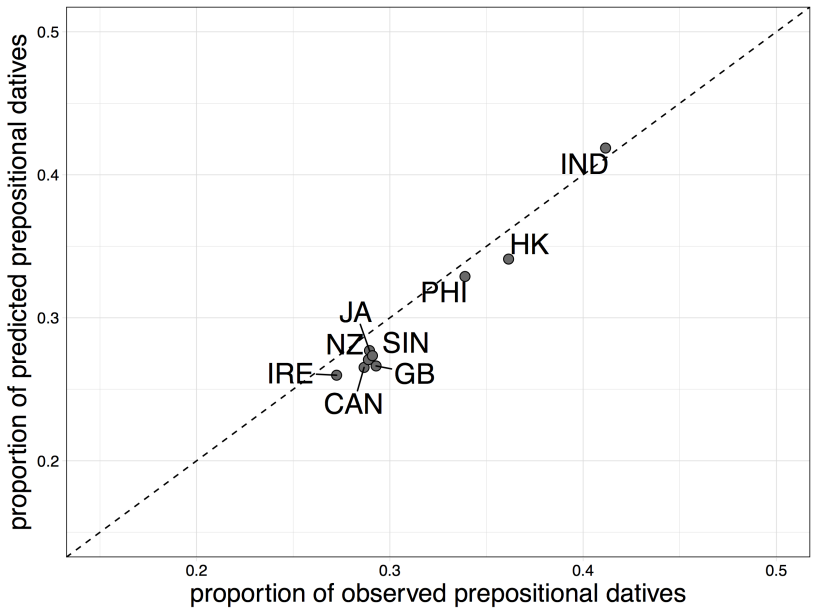


Figure 5.1 Proportion of observed versus predicted prepositional variant in all nine varieties — The predictions are best for IndE, in all other cases, the random forest underpredicts the use of the prepositional dative.

5.2.2 Relative importance of probabilistic constraints

Calculating the importance of predictors on a global scale (using the `varimpAUC()` function in the party package, Janitza et al. 2013) indicates that WEIGHTRATIO is the most important predictor for the choice of dative variant, closely followed by the pronominality of the recipient. Complexity of the theme, pronominality of the theme and theme head frequency follow after (see Figure 5.2). Note that VARIETY as well as MODE and CORPUS rank relatively low in importance.

These findings are consonant with most previous studies that also observed a dominant effect of relative length and recipient pronominality in their data (for instance, Schilk et al. 2013: 22; Bernaisch et al. 2014: 20; see, however, Bresnan et al. 2007a; Szmrecsanyi et al. 2017).

A comparison between the predictor rankings of the conditional random forests fitted separately by variety furthermore shows that length is consistently the most important factor followed by recipient pronominality, with the exception of Indian English (see Figure 5.3).

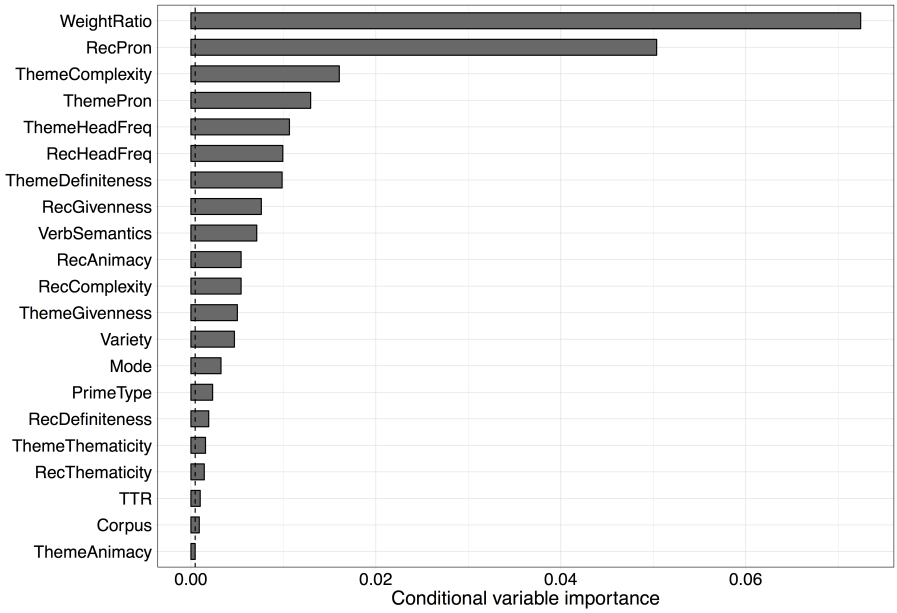


Figure 5.2 Variable importance of predictors fuelling variation in the dative alternation — The most important predictor is relative length, followed by recipient pronominality.

5.2.3 Interim summary

In sum, the conditional random forest points to the importance of relative length and recipient pronominality as decisive factors in the alternation between the ditransitive and the prepositional dative – a finding that is in line with previous work especially regarding the importance of recipient pronominality (e.g. Bresnan et al. 2007a; Schilk et al. 2013; Bernaisch et al. 2014) but deviating to some extent with regard to the importance of length (Szmrecsanyi et al. 2017). What is more, the random forest models fitted separately by variety indicate that IndE deviates from the other varieties in that recipient pronominality and not relative length constitutes the most important predictor.

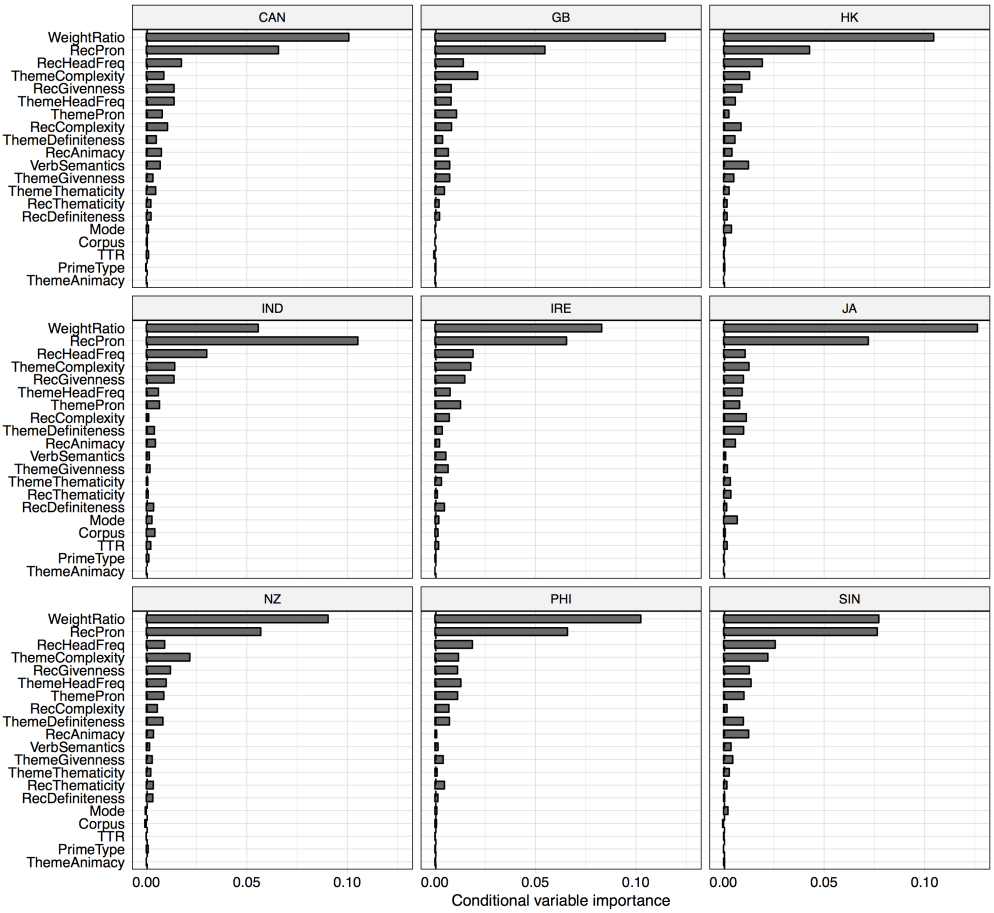


Figure 5.3 Predictor rankings by variety — Length is the most important predictor in all varieties with the exception of Indian English.

5.3 Probing the multivariate nature of dative choice

5.3.1 Model selection

To gauge regional variation in probabilistic constraints in the choice of dative variant in more detail, a mixed-effects logistic regression model was fitted to the data. In order to locate all possible cross-varietal contrasts, the initial model included all factors listed in Chapter 4 as fixed effects (apart from VERBSENSE). In order to identify possible interaction terms, two conditional inference trees were fitted to the dative data with the exact same model formula as (19) using the `ctree()` function from the partykit package (Hothorn & Zeileis 2015). If the same predictor was used in two nodes following a split, this was taken as indication of a possible interaction between the factor of the original split and the subsequent nodes. Different random

seeds were set for each tree in order to corroborate the tree's robustness. The initial model thus included higher order interactions of language-internal predictors with the language-external constraints VARIETY, MODE and CORPUS that were justified by the conditional inference tree. In addition, interaction terms were included if cross-tabulation of the predictor with any of the three language-external factors indicated a skewed distribution across the two dative variants. Other interactions were not considered since cross-tabulation of the data did not justify their inclusion and because convergence of the model failed. Numeric variables were scaled by two standard deviations and centred around the mean (Gelman 2008). VARIETY was coded using sum coding instead of the more frequently employed treatment coding in order to compare the proportion of responses for each level against the grand mean across all levels (see Menard 2010: 97) and not just against one reference level. The random structure included a random intercept for VERBSENSE nested into VERB, a random intercept for SPEAKERID nested into FILEID nested into GENREFINE nested into GENRECOARSE (to account for corpus structure), a random intercept for the lexical theme head and a random intercept for the lexical recipient head. Due to the sparseness and the abundance of hapax legomena of recipient and theme heads, infrequent recipients and theme heads were subsumed under a new level labelled 'OTHER'. The threshold of inclusion was set to 90%. In other words, 'OTHER' included those 90% of recipients that occurred less than 4 times (in the case of recipient head) and those 90% of all themes that occurred less than 8 times (in the case of theme head). Including the multiple levels of corpus structure as well as lexical-specific items in the random component is essential to ensure that the basic assumption of the non-independence of data points is not violated (Gries 2015: 99). Other random intercepts or slopes could not be considered due to failure in model convergence.

Model selection then followed the backward elimination process outlined by Zuur et al. (2009). Starting with the initial model, the maximum random structure was first identified by removing those random components that did not significantly improve the model fit according to likelihood tests. Next, the optimal fixed effects structure was determined in a similar process, first removing non-significant interaction terms, followed by non-significant main effects. The predicted outcome of the model is the log odds of the prepositional dative variant. The final model is given in (20). Note that three factors are included as interaction terms with VARIETY (see Section 5.3.4 on a discussion of the interaction terms).

(20) Variant ~ (1|VERB/VERBSENSE) + (1|GENRECOARSE/GENREFINE/FILEID) + (1|THEMEHEAD) + (1|RECHEAD) + THEMEANIMACY + RECANIMACY + RECCOMPLEXITY + THEMECOMPLEXITY + RECGIVENNESS + THEMEGIVENNESS + TYPETOKENRATIO + THEMEDEFINITENESS + RECDEFINITENESS + PRIMETYPE + WEIGHTRATIO + RECPRON + VARIETY + CORPUS + VARIETY:WEIGHTRATIO + VARIETY:RECPRON + VARIETY:CORPUS

Summary statistics for the model indicate that the model fits the data well. The model can predict 93.3% of the data accurately which is significantly better than the baseline of 73.84% (p_{binom}<.001). Somer's *C* index is an outstanding 0.98 and conditional and marginal R^2 values are 0.836 and 0.383 respectively (based on the `r.squaredGLMM()` function in the MuMIn package, Bartoń 2016; see also Nakagawa & Schielzeth 2013). Marginal R^2 indicates the variance in the data accounted for by the model with only fixed effects, conditional R^2 gives the variance accounted for by the model including both fixed and random effect structure. Comparison between the two R^2 values reveal that (0.383/0.836=) 45.8% of the variance accounted for by the model is due to lexical effects and corpus structure alone. Collinearity between the factors in the model was assessed with the condition number κ (following Belsley et al. 1980) and the variance inflation factors (VIF) (adapted from the rms package, Harrell 2016). Condition number κ equals 13.0 indicating medium collinearity (Baayen 2008: 182). The VIFs indicate that much of the estimated variance of higher order interactions with VARIETY is associated with the corresponding main effect.

To validate the model, the data was randomly divided 100 times into a training set (consisting of approximately 75% of the data, i.e. 9,862 observations) and a test set (consisting of approximately 25% of the data, i.e. 3,309 observations). Models were then iteratively fitted to the training set and the predictions were calculated on the corresponding test set, measuring the accuracy of each of these 100 models in the probability of correctly predicted outcomes. Mean accuracy was 90.6% which indicates a good model fit; the accuracy measures ranged from 89.3% for the poorest to 91.6% for the best fit. In what follows, I will first discuss the random effects of the model, followed by the main effects and finally, interaction terms.

5.3.2 Random effects

The random effects of VERB and THEMEHEAD account for the largest amount of variance in the random effect structure of the model (see Table 5.2). The importance

of these two lexical effects in the dative alternation has been reported elsewhere before (Bresnan & Ford 2010: 202) and will therefore be subjected to a more detailed exploration in what follows. While a closer look at recipient heads might also be argued for at this point, a detailed discussion of lexical considerations regarding the recipient is deferred to Section 5.5 since recipient heads add only minimally to this model. Table 5.2 further indicates that two effects of corpus structure (the nested effect of GENREFINE:GENRECOARSE and the effect of GENRECOARSE) play only a marginal role in the random effects structure, that is, there is not much variance in the data that can be ascribed to idiosyncrasies in the registers sampled. The texts sampled per register (by GENREFINE and GENRECOARSE) display some idiosyncrasy as is visible from the three-way interaction of FILEID:GENREFINE:GENRECOARSE but the interaction still accounts for less variance than verb- and theme-related random effects. Also note that there is not much variability across the different meanings of a verb (the effect of VERBSENSE:VERB), rather the lexical form of the verb itself bears the largest impact. Let us thus have a closer look at the effects of verbs and themes.

Table 5.2 Estimated variances and standard deviations of random effects in the model — Verb and theme head account for the largest amount of variance in the model.

Groups	**Variance**	**Standard deviation**
FileID:GenreFine:GenreCoarse	0.41902	0.6473
ThemeHead	1.59663	1.2636
RecHead	0.31968	0.5654
VerbSense:Verb	0.32559	0.5706
Verb	6.37160	2.5242
GenreFine:GenreCoarse	0.07094	0.2663
GenreCoarse	0.01312	0.1146

The verbs *explain, demonstrate* and *submit* show the highest preference for the prepositional dative (indicated by positive adjustments to the model's intercept), while the verbs *permit, wish* and *allow* display the highest preference for the ditransitive dative (indicated by negative adjustments to the model's intercept). These adjustments are visualised in Figure 5.4: The larger a verb's distance from the intercept (the dashed line), the stronger the structural preference for the prepositional (upper half of the figure) or the ditransitive dative (lower half of the figure). The intercept represents the mean (in case of sum coding) or reference level (in case of treatment coding) of all factors. Only verbs with adjustment values larger than ±2 are labelled.

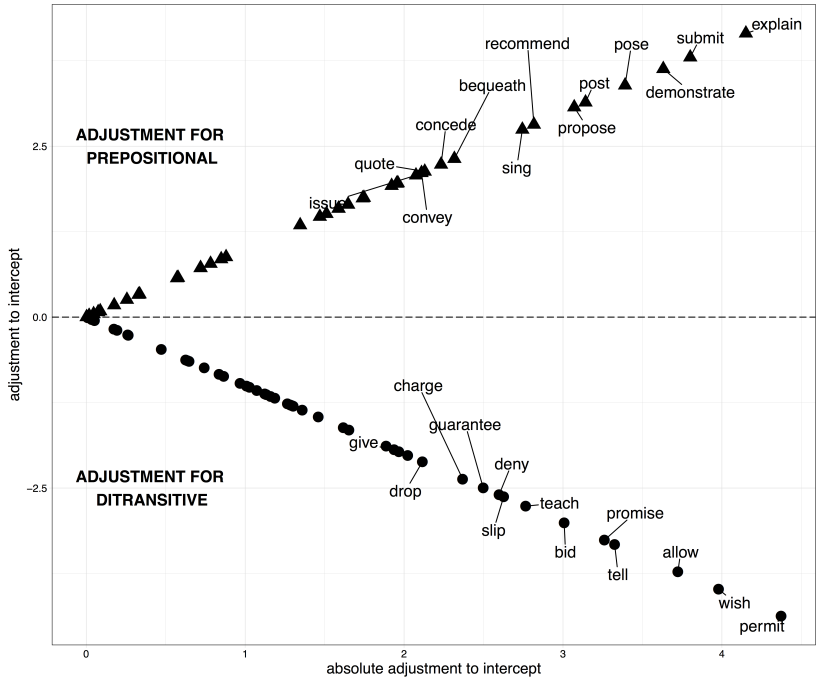


Figure 5.4 Constructional preferences by verb — Only verbs with an adjustment value larger than ±2 are labelled. Positive adjustments to the model's intercept (verbs in the upper half) signal preference for the prepositional dative, negative adjustments (verbs in the lower half) a preference for the ditransitive dative. The greater the absolute distance between the verb and the intercept (the dashed line), the larger the adjustment. The x-axis plots absolute adjustment values.

Results of the model's random structure are largely in line with previous research: Bresnan & Ford (2010: 178) find a bias towards the ditransitive dative for *allow* and *wish*, although in their data *tell* (in the communicative sense) and *pay* (in the abstract sense) show the largest adjustments to the model intercept with *tell* clearly favouring the ditransitive dative. One has to keep in mind, however, that their study samples only 38 verbs from spoken US English (the Switchboard corpus). Similarly, De Cuypere & Verbeke (2013) find a bias of *permit* towards the ditransitive dative and a bias of *submit* towards the prepositional dative in IndE, while Theijssen (2012: 17) finds a bias of *explain* towards the prepositional dative. Lexical preferences towards one or the other variant can override syntactic constraints, as is shown in the examples of *explain* (21) and *demonstrate* (22). Both verbs consistently favour the prepositional dative despite the possible influence of recipient pronominality and relative length on the choice of dative variant.

(21) *..., Mrs. Joan Lynn* ***explained some of the different elements of Jewish family life to us*** <ICE-IRE:W2D-020>

(22) *..., but also* ***demonstrates some interesting internet statistics to you,*** *...* <GloWbE-HK:B:3581709>

Note that *demonstrate* occurs eight times in the data and all eight times in the prepositional dative, whether the verb was used in the communicative or abstract sense. Similarly *explain* occurs 72 times in the data out of which only two times in the ditransitive pattern (all 72 instances in the communicative sense).

Regarding the themes, *try, go, choice* and *fee* show the highest adjustments for the ditransitive dative, while *it, birth, attention* and *them* display the highest adjustments towards the prepositional dative. Figure 5.5 plots positive adjustments above the dashed line (preference for prepositional dative) and negative adjustments below the dashed line (preference for ditransitive dative) and labels only those themes with an absolute adjustment value larger than 1.8 (for reasons of visualisation). The larger the adjustment, the longer the distance from the model's intercept (represented by the dashed line).

Themes with the highest adjustments towards the ditransitive dative include themes that often occur in very idiomatic expressions like in (23) and (24) as well as themes such as *choice* or *fee* that mainly occur in the ditransitive dative but variably with different verbs and recipients (see examples 25 and 26). Despite this lexical variability in the use of *choice* and *fee*, the latter still often co-occurs with the verb *pay* pointing to some sort of idiomaticity or at least lexically entrenched co-occurrence. In addition, apart from one exception each, all uses of *choice* and *fee* are used with verbs denoting abstract meaning (according to verb semantics). This goes hand in hand with previous findings that have found the abstract meaning of *give* to greatly prefer the ditransitive dative in contrast to *give* denoting transfer of concrete items (Röthlisberger et al. 2017). Both findings attest to the preferred use of abstract verb semantics in ditransitive datives. Note also that idiomatic or fixed expressions with *try* and *go* were retained in the dataset since their variability in those varieties was verified with Google or GloWbE (as discussed in Chapter 4, Section 4.2.3).

(23) *... and finally he agreed to* ***give it another try***. <GloWbE-IND:B:3468930>

(24) ***Give it a go***. <GloWbE-SIN:B:3524241>

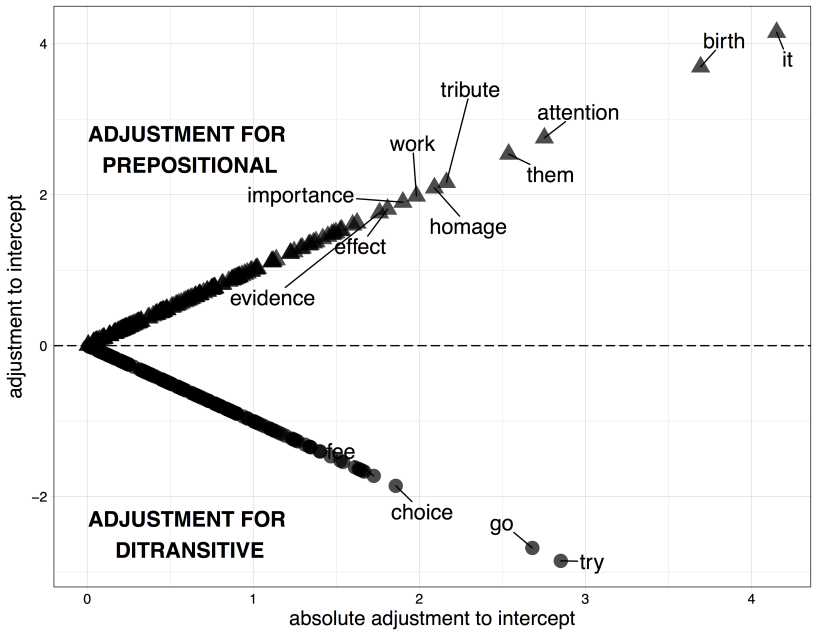


Figure 5.5 Constructional preferences by theme — Positive adjustments to the model's intercept (themes in the upper half) indicate a preference for the prepositional dative, negative adjustments (themes in the lower half) a preference for the ditransitive dative. The greater the distance between the theme and the zero-intercept (the dashed line), the larger the adjustment. Only themes with an adjustment value above ±1.8 are labelled.

(25) *For example: We can't afford to concede any more but we are prepared to* ***offer you a choice****.* <ICE-NZ:W2D-011>

(26) *Also you can also* ***pay them a successful fee*** *if there is a deal close from their referral.* <GloWbE-SIN:G:1018875>

The list of themes that strongly prefer the prepositional dative is led by *it*, *birth*, *attention* and *them*, thus including themes that occur in seemingly fixed expressions as in (27) and (28). These observations were retained in the data because verification with Google or GloWbE attested their use in both dative variants. The high preference for prepositional datives with *pay* co-occurring with *attention* is confirmed. The high preference of *it* and *them* for the prepositional dative is not surprising, given the fact that both are pronominal and short constituents (see example 29). As previously shown, short and pronominal constituents tend to be expressed early when speakers

have a word order choice (Hawkins 1994). Nevertheless, both *it* and *them* can also occur in the ditransitive as shown in (30).

(27) *Love of the transcendental image of the Virgin Mary was what* ***gave birth to the glorious cathedrals of the Gothic world****.* <ICE-GB:W1A-008>

(28) *Greece* ***paid no attention to it****.* <GlowbE-GB:G:387816>

(29) *And then as soon as I got the first roll* ***give it to you*** *and then roll it up.* <ICE-CAN:S1A-045>

(30) *No man, give, just* ***give me them****.* <ICE-JA:S1B-007>

The adjustments to the model's intercept only allow for a global perspective aggregated over all nine varieties. This global perspective offers some insights into the lexical preferences of verbs and themes as a whole but provides no information on regional variation with regard to the constructional preferences of the variants' lexical items. Such regional variation will be explored in Section 5.5.

5.3.3 Main effects

The coefficient estimates of the main effects in the model are summarised in Table 5.3. Estimates of the coefficients are given on a logit-scale in the column labelled β. Positive values indicate a preference for the predicted outcome (the prepositional dative), and negative values indicate a preference for the ditransitive dative. *SE* specifies standard errors.

While the coefficient estimates of all main effects are given in Table 5.3, only the value of those coefficient estimates should be taken at face value where the factor itself does not form part of an interaction (see Crawford et al. 2014 and Levy 2014). The effect of all other main effects that constitute an interaction term on the choice of dative variant is illustrated by way of univariate mosaic plots (Figure 5.6). Mosaic plots visualise the distribution of frequency and raw numbers in contingency tables.

The results of the main effects can be summarised as follows: First, the constraints in the model have the expected effect given the literature. For instance, if the recipient is pronominal, the ditransitive is more likely, as in (31a) and if the recipient is non-pronominal, the prepositional dative is more likely, as in (31b). Weight ratio also

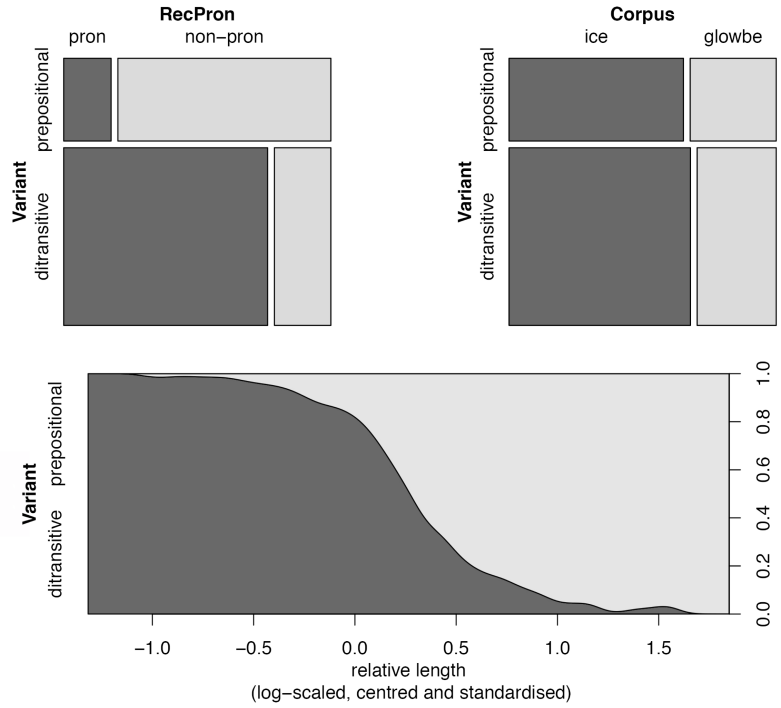


Figure 5.6 Univariate mosaic plots of the main effects that also form part of an interaction in the mixed-effects model — Each figure shows the proportional distribution of the two dative variants across the levels of the predictor.

has the expected effect in that the more the recipient increases in length in relation to the theme, the greater the odds for a prepositional dative. Similarly, the effects of animacy, givenness, definiteness and complexity are congruent with the findings of previous research: Whenever a constituent is given, animate, definite or simple, the model indicates that language users tend to place it first in the ordering of the constituents (exemplified in 32 and 33). In other words, if the recipient is given, animate, definite or simple, the ditransitive is the preferred option. If the theme is given, animate, definite or simple, the prepositional dative increases in likelihood. Also, priming or structural persistence has the predicted effect given the literature: The previous occurrence of a ditransitive dative increases the likelihood of another ditransitive dative; the previous occurrence of a prepositional dative increases the likelihood of a prepositional dative. Note that the priming variant occurs in the ten utterances preceding the dative token and that only previous alternating variants were considered for priming effects. Furthermore, the ditransitive dative is also more likely if type-token ratio increases, that is, if the lexical density of the 100 words

surrounding the dative token increases.

(31) a. *The Data Protection Act* ***gives you this right***. <ICE-GB:W2D-010>

b. *And he* ***gave her to Basil***. <ICE-IRE:S1B-003>

(32) a. *Video reruns hinted that Dowie had impeded a defender thus* ***giving Quinn the necessary space***. <ICE-IRE:W2C-001>

b. *we* ***assign our students to certain tutorial groups***. <ICE-SIN:S2A-047>

(33) a. *I mean I can say that he* ***caused my property physical damage***. <ICE-NZ:S1B-012>

b. *Last month, I agreed to* ***sell my computer to a friend***. <ICE-SIN:W2B-002>

Second, the likelihood of a prepositional dative vis-à-vis a ditransitive dative is generally the same across all varieties (as a main effect) with four exceptions: Speakers of Indian and Hong Kong English as well as Irish and Canadian English deviate from the global average. In India and Hong Kong, the likelihood of a prepositional dative is statistically significantly higher than in the rest of the varieties; in Irish and Canadian English, the likelihood of a ditransitive is higher than in the rest of the varieties. Note that this regional difference is observable when all other factors are at their reference level.

Table 5.3 Main effects of individual factors in the model — Model predictions are for the prepositional dative. Only significant factors shown.

Factor	β	***SE***	***p***
(Intercept)	-1.957	0.421	<0.001
THEMEANIMACY: inanimate ⇒ animate	0.875	0.342	0.011
RECANIMACY: animate ⇒ inanimate	0.840	0.113	<0.001
THEMEBINCOMPLEXITY: complex ⇒ simple	0.843	0.122	<0.001
RECBINCOMPLEXITY: simple ⇒ complex	1.084	0.158	<0.001
THEMEGIVENNESS: new ⇒ given	0.265	0.098	0.007
RECGIVENNESS: given ⇒ new	0.307	0.098	0.002
THEMEDEFINITENESS: indef ⇒ def	0.640	0.098	<0.001
RECDEFINITENESS: def ⇒ indef	0.580	0.108	<0.001
RECPRON: pron ⇒ non-pron	1.656	0.231	<0.001
WEIGHTRATIO (log)	2.783	0.177	<0.001
TTR	-0.2561	0.086	0.003
CORPUS: ice ⇒ glowbe	-0.247	0.263	0.346
PRIMETYPE			
none ⇒ do	-0.281	0.115	0.015
none ⇒ pd	0.433	0.143	0.002
VARIETY			
ALL ⇒ CAN	-0.759	0.238	0.001
ALL ⇒ HK	0.699	0.170	<0.001
ALL ⇒ IND	0.808	0.187	<0.001
ALL ⇒ IRE	-0.479	0.231	0.038

5.3.4 Interaction terms

Of main interest for the study are the interaction terms in the model since those indicate the contexts in which the probabilistic constraints on dative choice are regionally variable in their effect size (and direction). Interaction terms test the effect of the levels of one factor against the levels of another factor – say, for instance, the effect of animate and inanimate recipients against the nine different levels of variety. The assumption behind interaction terms is that the levels of the first factor (e.g. recipient animacy) are believed to behave differently depending on the levels of the other factor (e.g. VARIETY). Since this study is interested in regional variation, interaction terms with VARIETY were included in order to test whether the effect of any of the other factors was significantly different between varieties. After backwards elimination, only three factors remained as significant interaction terms in the model (see Table 5.4) – that is, three factors turn out to be regionally malleable, namely weight ratio (length), recipient pronominality and corpus. Let me elaborate on each of these interactions briefly.

Table 5.4 Interaction effects in the model between VARIETY and language-internal factors and CORPUS — Model predictions are for the prepositional dative (only significant factors shown).

Factor	β	***SE***	***p***
VARIETY : WEIGHT RATIO			
IndE	-0.842	0.338	0.013
IrE	-0.724	0.339	0.033
JamE	1.076	0.414	0.009
VARIETY : RECPRON			
HKE + non-pron	-0.656	0.241	0.007
IndE + non-pron	0.904	0.264	<0.001
VARIETY : CORPUS			
IndE + glowbe	-0.914	0.243	<0.001

The interaction effect between VARIETY and WEIGHTRATIO is smaller in Indian and Irish English, the two varieties where the likelihood of the prepositional does not increase as much as in the other varieties when the recipient increases in length compared to the theme. The effect is strongest in Jamaican English compared to all

other varieties, where the prepositional dative becomes even more likely when the relative length of recipient and theme increases. The increase in the likelihood of a prepositional dative with increasing relative length is illustrated in Figure 5.7 which visualises this effect in all nine varieties. Note that the cline is steeper, indicating a stronger effect, in Jamaican English and flatter, indicating a weaker effect, in Indian and Irish English compared to the global average (the three varieties are positioned in the middle row and highlighted in grey). Effects are plotted with the effects package (Fox 2003).

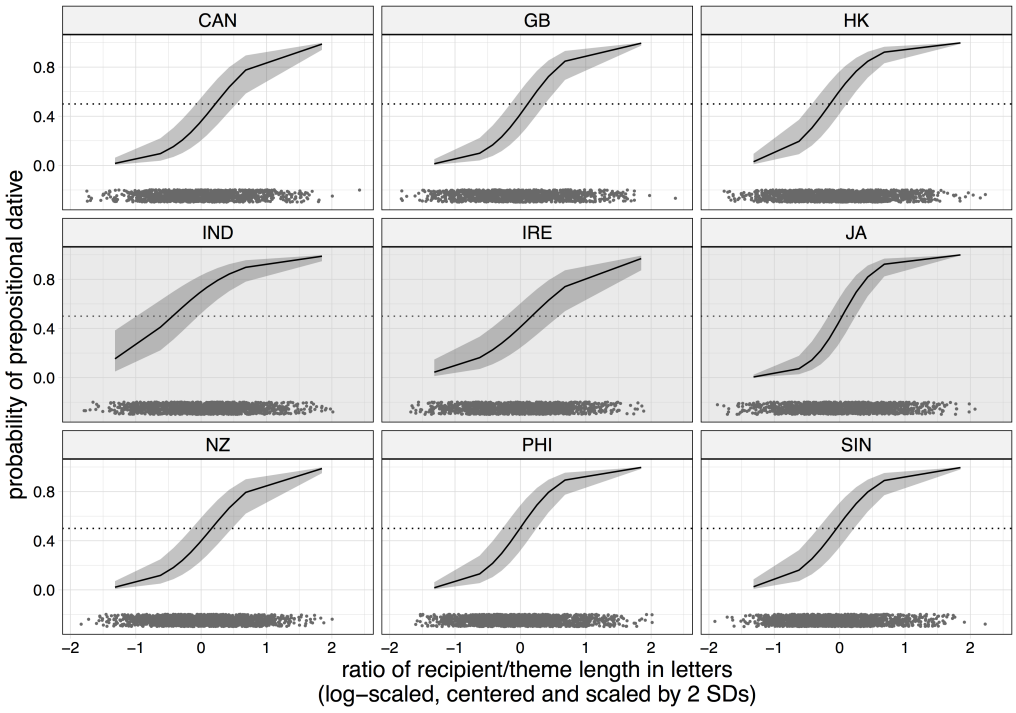


Figure 5.7 Effect of weight ratio by variety — The prepositional dative becomes more likely in Jamaican English and less so in Indian and Irish English when WEIGHTRATIO increases (varieties highlighted in grey), compared to the global average.

The interaction effect between VARIETY and RECPRON is significantly different in Hong Kong and Indian English compared to the global average. In Hong Kong English, a non-pronominal recipient does not increase the likelihood of a prepositional dative as much as in the other varieties. In Indian English, on the other hand, the effect of a non-pronominal recipient is stronger and the prepositional dative more likely compared to all other varieties. This stronger effect is reflected in the wide gap in Indian English in Figure 5.8 between the effect of pronominal (solid line) and non-pronominal

recipients (dashed line) on the likelihood of a prepositional dative. Figure 5.8 plots the likelihood of a prepositional dative if the recipient is pronominal (solid line with solid circles) and non-pronominal (dashed line with triangles) per variety. Besides the large difference in likelihood between pronominal and non-pronominal recipients in Indian English, the small difference in likelihood in Hong Kong English is also statistically significantly different from the global average.

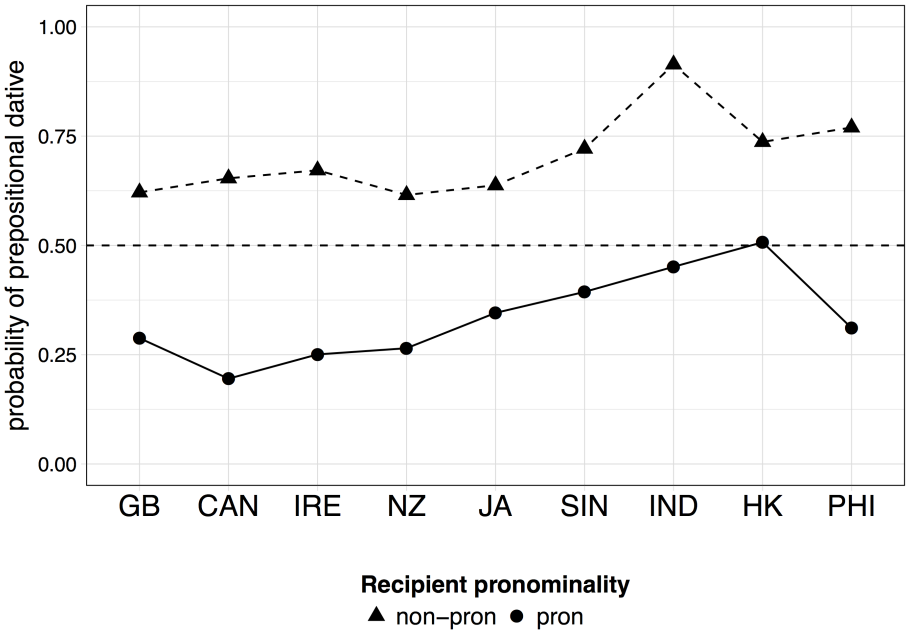


Figure 5.8 Effect of recipient pronominality by variety — The likelihood of a prepositional dative increases more in Indian English and less in Hong Kong English when the recipient is non-pronominal instead of pronominal, compared to the global average. Native varieties appear on the left side, non-native varieties appear on the right side of the graph.

The effect of the interaction between VARIETY and CORPUS is only significant in Indian English. The coefficient estimates indicate that the proportional distribution of ditransitive and prepositional datives in the two corpora used for this study, ICE and GloWbE, is statistically significantly different in Indian English: Prepositional datives are more frequent in ICE compared to GloWbE (see also Figure 4.2 in Section 4.3). This difference is reflected by the large distance between the probability of prepositional datives in ICE (dashed line) and the probability of prepositional datives in GloWbE (solid line) in Indian English (IND) in Figure 5.9. The difference between ICE-IND and GloWbE-IND might be due to the different sampling periods of ICE (1990s) and

GloWbE (early 2010s) or due to an effect of register. Since GloWbE samples only one specific register – namely online blogs and websites – a closer look at register effects is certainly warranted (as done in Section 5.6).

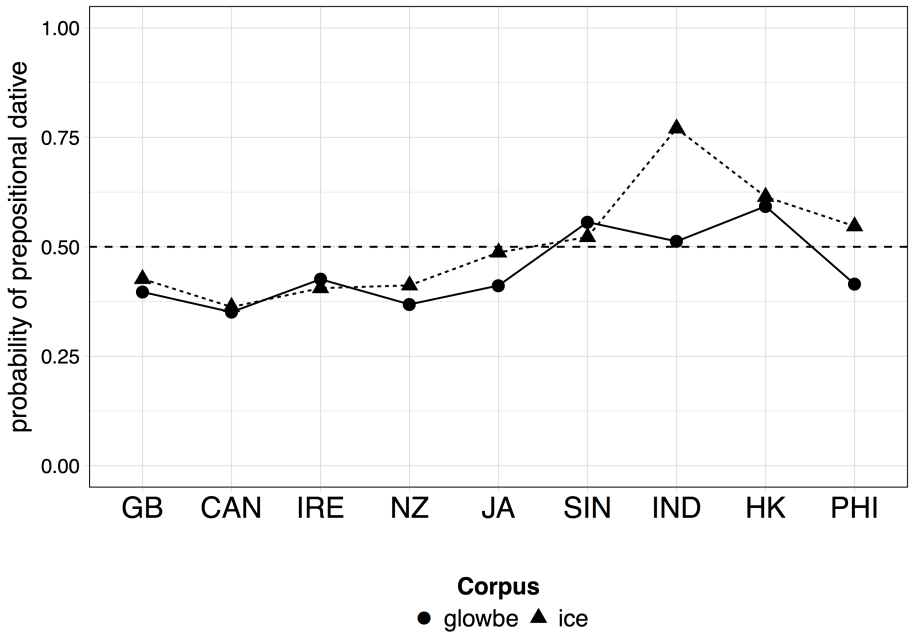


Figure 5.9 Effect of corpus by variety — The likelihood of a prepositional dative is higher in Indian English if the data stems from ICE instead of GloWbE, compared to the global average. Native varieties appear on the left side, non-native varieties appear on the right side of the graph.

The statistically significant interactions with recipient pronominality and with weight by variety were also observed in an earlier analysis of the data when attention was restricted to ICE only (Röthlisberger et al. 2017).

5.3.5 Interim summary

The multivariate analysis of the dative data using mixed-effects logistic regression reveals that the factors influence the choice of dative variant as predicted given the literature. The influence of these factors is thereby congruent: Speakers tend to opt for that dative variant where the first constituent is easier to process and produce than the second one – easier meaning animate, definite, pronominal, short(er) and so on. While the effect **direction** is thus constant cross-regionally, the effect **size** of three constraints, namely WEIGHTRATIO, RECPRON and CORPUS, differs from the global

average in four varieties, that is, in Indian, Hong Kong, Irish and Jamaican English, to varying degrees (summarised in Table 5.5): The relative length of recipient and theme has a stronger effect in Jamaican English and a weaker effect in Indian and Irish English compared to the global mean. Recipient pronominality has a stronger effect in Indian English and a weaker effect in Hong Kong English compared to the global mean in that speakers of Indian English are even more likely to use the prepositional dative if the recipient is nominal instead of pronominal. Finally, only in IndE does CORPUS have a statistically significant weaker effect than in the eight other varieties: In IndE, speakers are less likely to use a prepositional dative in online blogging and websites than in ICE compared to all other varieties.

Table 5.5 Cross-varietal differences in effect size — Minus (-) indicates decreased effect size, plus (+) indicates increased effect size compared to the global mean.

Variety	WEIGHTRATIO	RECPRON	CORPUS
IrE	−	=	=
IndE	−	+	−
JamE	+	=	=
HKE	=	−	=

The malleability of weight ratio and recipient pronominality as well as of CORPUS warrants further exploration. Hence, the next three sections each pay closer attention to the cross-lectal variability of these three predictors. Section 5.4 zooms in on end-weight effects and supplements length measurements with another fine-grained measure to gauge syntactic complexity besides the binary predictor introduced previously. Section 5.5 focuses on recipient pronominality and aims to assess regional differences in the lexical items that instantiate recipients. Since lexical effects might well play a role in the choice of dative variant (see, for instance, Gries & Stefanowitsch 2004), the analysis is extended to also include themes and verbs. Section 5.6, then, takes the cross-varietal differences in the effect of CORPUS as a starting point to probe the register-specificity of the English dative alternation further.

5.4 Regional variation of end-weight effects

Until now, this study has made use of constituent length as a proxy to gauge the effect of end-weight on dative choice. Using length as a proxy of end-weight is not

completely unwarranted since previous work has illustrated that the length of a constituent highly correlates with other measures of end-weight (e.g. the number of syntactic nodes) (see Wasow 1997b, 2002; Szmrecsanyi 2004; Shih & Grafmiller 2011) and can thus serve as a near-proxy for end-weight in general (for instance, Szmrecsanyi 2004; Rosenbach 2005; Bresnan et al. 2007a) since structurally heavy constituents tend to be long, structurally light constituents tend to be short (Berlage 2014: 7). Even though the correlation between different measures of end-weight is generally high, some studies have also shown that these measures can affect linguistic variation independently of each other. The focus of these studies has thereby been on pitting different weight measures against each other in order to explore the nature of end-weight and the correlation between the various measurements more closely. For instance, in an experimental setting, Ferreira (1991) tests the initiation times for utterances that share the same length but differ in their degree of structural complexity (i.e. utterances without postmodification vs. utterances with prepositional phrases as postmodifier vs. utterances with sentential postmodification). Despite the apparent strong correlation between length and structural complexity, structural complexity of the noun phrase is shown to affect processing independent of length. Shih & Grafmiller (2011) probe the influence of various weight-related measures (e.g. number of nodes, words, syllables, stressed syllables, phonemes) on different syntactic alternations. Their results reveal that the number of nodes is the most important predictor on genitive choice, while the length of constituents is the most influential factor on dative choice, indicating that the importance of the various constraints depends on the variable that is investigated. Similarly, Berlage (2014: 250-251) concludes her book-length treatment of noun phrase complexity by suggesting that length of the noun phrase is of major importance in word order alternations (such as the genitive or dative alternation) while structural complexity of the noun phrase is crucial in lexical variation. Since Berlage uses a novel annotation scheme that takes the *nouniness* of the post-head dependents into account (as defined by Ross 2004[1973]) and makes testable predictions about the dative alternation, I will use her study as a starting point to analyse the interplay between length measurements and syntactic/structural complexity in the dative alternation more closely. As will be shown, her predictions regarding the effect of length versus syntactic complexity in word order alternations find support in the current data.

Besides this interest in the various measures that gauge end-weight, attention has recently shifted to the variability of end-weight effects across registers (see Grafmiller

2014: 484-485), varieties (e.g. Bresnan & Ford 2010) and across time (e.g. Wolk et al. 2013). No attention has so far been paid to the extent to which different end-weight measures might be differently malleable cross-lectally, the assumption being that the high correlation between the various measures guarantees the same effect for all of them. As this section will illustrate, not all factors gauging end-weight effects are in fact amenable to cross-lectal variability, at least not in the case of the dative alternation.

Following Berlage (2014), the current section will thus add structural complexity to the analysis and leave aside other possible ways to gauge end-weight effects (see Shih & Grafmiller 2011). While Berlage does not explicitly focus on the dative alternation in her work on noun phrase complexity, she nevertheless (tentatively) proposes, based on her case studies, that

> cases of word-order variation behave differently from those that operate with the optional occurrence of a syntactic item. If this were the case, we may expect such variables as the genitive and the dative alternation to be more sensitive to the length of the NP in question (e.g. the possessor/ possessum in the genitive variation and the theme or the goal in the dative alternation) than to the type of postmodifier present. (Berlage 2014: 251)

The aim of the current section is thus on the one hand to expand on Berlage's study by (1) determining the importance of structural complexity and length on dative choice separately and thereby address the question whether length is indeed more important than structural complexity, and – in view of the cross-regional malleability of length effects in the dative alternation – (2) to investigate the extent to which structural complexity is also variable in its effect size across regional varieties of English. In addition, the analyses will (3) be repeated on a dataset restricted to nominal constituents only. By delimiting the variable context so restrictively, I can investigate the extent to which length and structural complexity are regionally variable if highly correlated constituents such as pronouns are excluded. In light of the findings from these three analyses, the final part of this section will address the question whether a five-level predictor of structural complexity is really necessary.

Structural complexity is defined in this study both in absolute-quantitative as well as qualitative terms: Following Berlage (2014) and others, the employed measure of structural complexity takes both the number of post-head dependents as well as the *nouniness* of each constituent's post-head dependents into account (see Section 5.4.1 for detailed elaboration on the coding procedure). Regarding the definition

of *nouniness*, I rely on Ross (2004[1973]) who proposes a scale of least to most noun-y constructions based on a series of syntactic test frames (such as preposition deletion, extraposition, pied piping and so on). He exemplifies this scale by ranging various types of complements (for instance, embedded questions and *that*-clauses) from being very noun-like to having sentence-like properties depending on the construction's ability to undergo the syntactic test frames. Since each dative token contains two constituents, their structural complexity is calculated both separately by constituent and as a ratio.

The present section first introduces the coding of structural complexity (Section 5.4.1) before investigating the interdependence between and independent importance of complexity and length on a global as well as local level in more detail (Section 5.4.2). Section 5.4.3 presents the results of analyses that gauge the cross-regional malleability of structural complexity in the dative alternation. Cross-regional malleability is further explored in Section 5.4.4 which investigates the extent to which length and structural complexity are regionally variable in their effect sizes if pronominal constituents are excluded. Finally, in light of the results, the last part of this thematic section follows up on the question whether a five-level predictor of complexity is really necessary.

5.4.1 Coding for complexity

The coding of structural complexity closely follows the methodology outlined in Berlage (2014). Coding was done fully manually and restricted to the ICE dataset, which was considered sufficiently representative to start with ($N = 9{,}058$).

Berlage (2014) distinguishes between two predictors, namely NP-length and NP-structure. NP-length corresponds to the factors already included in this study so far, that is the length of each constituent measured in the number of letters (RECLETTERLTH, THEMELETTERLTH) or words (RECWORDLTH, THEMEWORDLTH). NP-structure gauges both the number of post-head dependents and takes the nouniness of these dependent(s) into account (adapted from Ross 2004[1973]). I follow Huddleston & Pullum (2002: 329) in calling linguistic elements following a constituent's head *post-head dependents*, a term that refers to both modifiers and complements (Huddleston & Pullum 2002: 331). In order to account for constituents' nouniness when determining their degree of structural complexity, I rely on an adaption of Ross (2004[1973]) (see Berlage 2014: 14-18). Since all post-head dependents in my own data are headed

by nouns – and do not, as in Ross's case, constitute independent complements (e.g. *that*-clauses as in *I regret that you left*) – I essentially had to adapt Ross's scale of nouniness to the post-head dependents in the dative alternation. In accordance with Berlage's (2014) proposed complexity scale, which in turn relies on the results of empirical work by Wasow (2002) and Wasow & Arnold (2005) among others, I distinguish between post-head dependents containing verb phrases (= sentential) and those that do not (= nominal) (see Berlage 2014: 17).

The first round of annotations led to 16 different categories – from the most simple ('s') constituent to constituents with sentential post-head dependents, for instance adverbial clauses ('advc'), nominal clauses ('nc') or complement clauses ('cp') (a complete list is provided in Appendix A). Due to the sparsity of tokens in some of these levels and because this study also aims to take the number and not just the type of post-head dependents into account, these 16 levels were further conflated to create a five-level predictor of structural complexity, namely RECCOMPLEXITY5 and THEMECOMPLEXITY5 (see Table 5.6). The five levels were defined in such a way as to offer a possible scale of increased complexity. This proposed scale of increased complexity is based on the assumptions that, first of all, constituents with no post-head dependents are less complex than constituents with nominal post-head dependents which are in turn less complex than constituents with sentential post-head dependents. And second, constituents with one post-head dependent are less complex than those with two or more post-head dependents (see Berlage 2014: 15, 58-65). The first level thus includes all simple constituents without post-head dependents, the second level includes all constituents with one post-nominal (i.e. non-sentential) dependent, the third level encompasses constituents with one sentential post-head dependent, the fourth level includes constituents with two or more nominal post-head dependents and the last level includes constituents with two or more post-head dependents of which at least one is sentential (see Table 5.6).

The proportional distribution of dative variants across the five levels of RECCOMPLEXITY5 and THEMECOMPLEXITY5 highlights that simple recipients are mainly expressed in the ditransitive dative (dark grey bar, left bar plot in Figure 5.10). Figure 5.10 further shows that recipients with post-head dependents are more often expressed in the prepositional dative where they are positioned last. To the extent that speakers are more likely to choose a prepositional dative when recipient complexity increases, the distributions shown in the bar plot support the proposed complexity scale as follows: Since the proportional distribution indicates that prepositional

Table 5.6 The five levels of NP-structure in the dative data — The levels are listed according to their proposed scale of complexity from top (least complex) to bottom (most complex).

Code	Category	Examples
's'	simple constituents without post-head dependents	*you, subscriptions, any old rubbish*
'spp'	constituents with one post-head dependent such as a prepositional phrase, conjuncts, a general extender, *s*-genitives, a nominal adposition or a post-nominal adjective, adverb or determiner	*the lies about Obama, my father's gun*
'svp'	constituents with one sentential post-head dependent, e.g. non-finite, relative, complement, nominal or adverbial clause	*the guy that cause the accident, people injured on the street*
'mpp'	constituents with two or more nominal post-head dependents	*somebody here in Singapore, the officer in charge of foreign students*
'mvp'	constituents with two or more post-head dependents of which at least one is sentential	*a very good idea of how two out of the three temples at Paestum are actually laid out in relation to the rest of the city*

datives are more likely if the recipient is followed by a sentential ('svp') instead of a nominal ('spp') post-head dependent, we can assume that sentential post-head dependents are more complex than nominal ones. Prepositional datives are also more likely if the recipient is followed by more than one post-head dependent ('m—' vs. 's—') supporting the hypothesis that multiple post-head dependents are more complex than just one post-head dependent. However, the bar plot on recipient complexity highlights one discrepancy in the proposed scale of complexity, namely that recipients with multiple sentential post-head dependents ('mvp') are less often expressed in a prepositional dative than recipients with multiple nominal post-head dependents ('mpp').

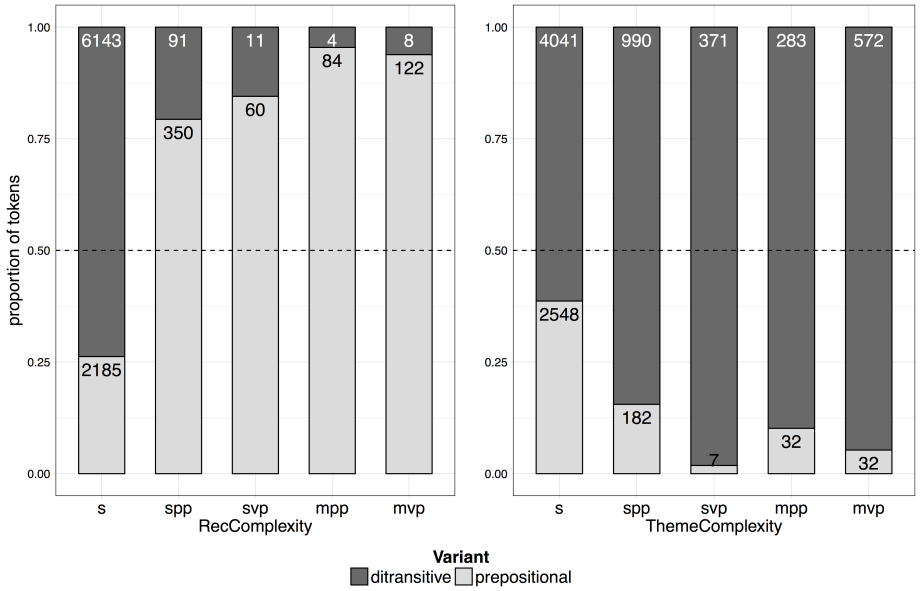


Figure 5.10 Proportional distribution of ditransitive and prepositional variants across the five levels of RECCOMPLEXITY5 (left) and THEMECOMPLEXITY5 (right) — Both bar plots show an overall increase in the proportion of that variant in which the last constituent is gaining in structural complexity following the proposed complexity scale.

Turning to theme complexity, Figure 5.10 (bar plot on the right) shows that simple themes ('s') are most often expressed in the ditransitive dative, similarly to recipients. If a post-head dependent is present, the frequency of ditransitive datives increases. If we again assume that a higher use of ditransitive datives with themes indicates an increase in complexity (since the theme is positioned last in the ditransitive dative), we can interpret the bar plot as follows: Simple themes ('s') are the least complex of all themes. Themes with a sentential post-head dependent ('svp') are more complex than themes with a nominal post-head dependent ('spp') as the proportion of ditransitive datives is higher for themes with a sentential post-head dependent. The same scale can be observed with regard to multiple post-head dependents. However, the figure again shows a discrepancy in the proposed scale of complexity: Themes with one sentential post-head dependent ('svp') have the highest proportion of ditransitive datives of all five complexity levels and are thus seemingly more complex than themes with multiple post-head dependents ('m—').

The overall pattern that emerges partly confirms the assumption that more complex constituents increase the likelihood of that dative variant where the complex constituent is expressed last. Overall, constituents with sentential post-head depen-

dents seem to be more complex than those with nominal post-head dependents, constituents with multiple post-head dependents seem to be more complex than those with one single post-head dependent. The discrepancies pointed out above, however, call into question the proposed scale of complexity. That the five levels of complexity might not evolve in a linear fashion from least to most complex is indicative that either a constituent's complexity level cannot be derived from its position in a dative variant or that a ratio rather than separate complexity values are needed.

Similar distributions of variants by complexity level can also be observed locally by variety. In each variety, speakers opt more frequently for the prepositional dative if the recipient is followed by sentential post-head dependents rather than nominal ones and if the number of post-head dependents increases (see Figure 5.11). Similarly, speakers of all varieties use the ditransitive dative more often if the theme is followed by sentential post-head dependents instead of nominal ones and if the number of post-head dependents increases (see Figure 5.12). Note that in Indian English, the prepositional dative is in fact the preferred option with simple recipients, in contrast to all other varieties where the ditransitive dative is more frequent. The proportion of prepositional datives with simple themes is also comparatively high in Hong Kong and Philippine English compared to the other varieties.

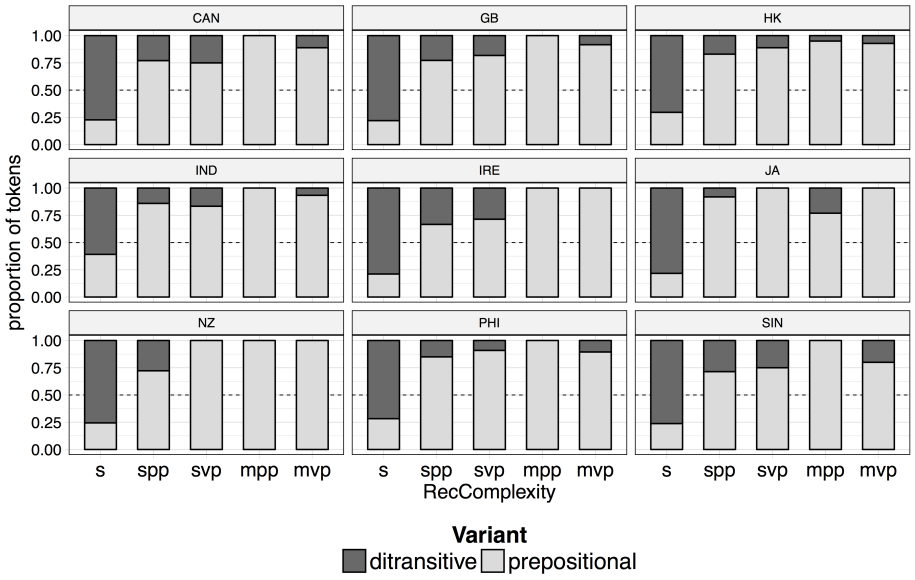


Figure 5.11 Proportional distribution of dative variants across the five levels of REC-COMPLEXITY5 per variety

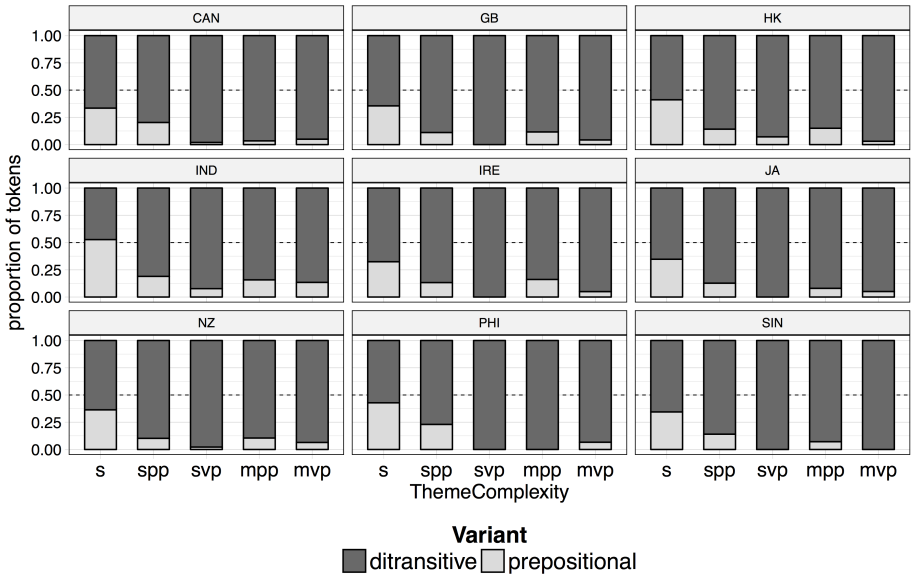


Figure 5.12 Proportional distribution by variety of dative variants across the five levels of THEMECOMPLEXITY5 per variety

5.4.2 The inter- and independence of complexity and length

Next, the correlation between NP-length and NP-structure and the predictors' individual importance regarding the choice of dative variant on a global as well as local level was assessed. That NP-length and NP-structure are highly correlated is not surprising given the fact that complex constituents tend to be long and simple constituents tend to be short (Wasow 1997b, 2002). A wide range of studies have pointed out, however, that structural complexity can influence syntactic variation independently of length and vice versa. In order to investigate the effect of NP-length and NP-structure more closely, their interdependence as well as independence of each other will take centre stage in what follows.

To gauge the correlation between NP-length and NP-structure, the levels of NP-structure were transformed into a numeric (ordinal) scale with 1 indicating simple constituents and 5 indicating constituents with multiple (sentential) post-head dependents, thereby following the proposed complexity scale. Figure 5.13 visualises the average correlation of NP-length (y-axis) across all five levels of complexity (x-axis) with a smoothed regression line plotted separately by recipients (solid line) and by themes (dashed line). Individual data points from that correlation are included separately for recipients (triangles) and themes (circles). Across all varieties, simple

recipients are short and complex recipients are long as indicated by the upward solid line in Figure 5.13. Similarly, simple themes are short and complex themes are long, which is reflected by the upward dashed line in Figure 5.13. The comparison between the variants reveals that simple recipients are always shorter than simple themes but complex recipients are overall longer than complex themes, with the exception of Irish English. Pearson product-moment-correlation coefficients (r) calculated between NP-length and (numeric) NP-structure (using `cor.test` in R) further reveal that the complexity of the theme correlates better with the theme's length ($r(9056) = 0.792$, $p<.001$) than the complexity of the recipient with the recipient's length ($r(9056) = 0.748$, $p<.001$).

This asymmetry between the constituents might be due to the fact that NP-length factors in premodifications while NP-structure does not. That is, NP-length distinguishes between constituents that are premodified and those that are not, irrespective of whether a constituent is followed by post-head dependents. For instance, *the young man* is longer than *the girl* in the number of letters but both NPs constitute simple noun phrases without any post-head dependents. Hence, premodified constituents that are not followed by post-head dependents can be either long or short but are, in any case, always 'simple'. Constituents with post-head dependents are necessarily long since every post-head dependent adds to the constituent's length. What is more, constituents with post-head dependents are not necessarily longer than simple constituents since simple constituents can increase in length with more premodifications as well.

The lower correlation between recipient length and recipient complexity compared to theme length and theme complexity might therefore be due to long but simple recipients. In other words, recipients are not necessarily complex but still as long as complex themes, while an increase in theme length is more often connected with an increase in complexity.

A comparison across simple constituents only supports this view: The longest simple recipient is a full 82 letters long while the longest simple theme is only 48 letters long. Taking a closer look at the 0.5% longest simple themes and recipients further reveals 36 recipients ranging from 33 to 82 letters in length and only 27 themes ranging from 32 to 48 letters in length (see examples 34a and 34b). 17 out of the 36 long but simple recipients turn out to be adpositions where the modifying elements precede the noun as in (34a). The rest of the mostly animate recipients are company names or premodified with more than one adjective or adverb (see 34b).

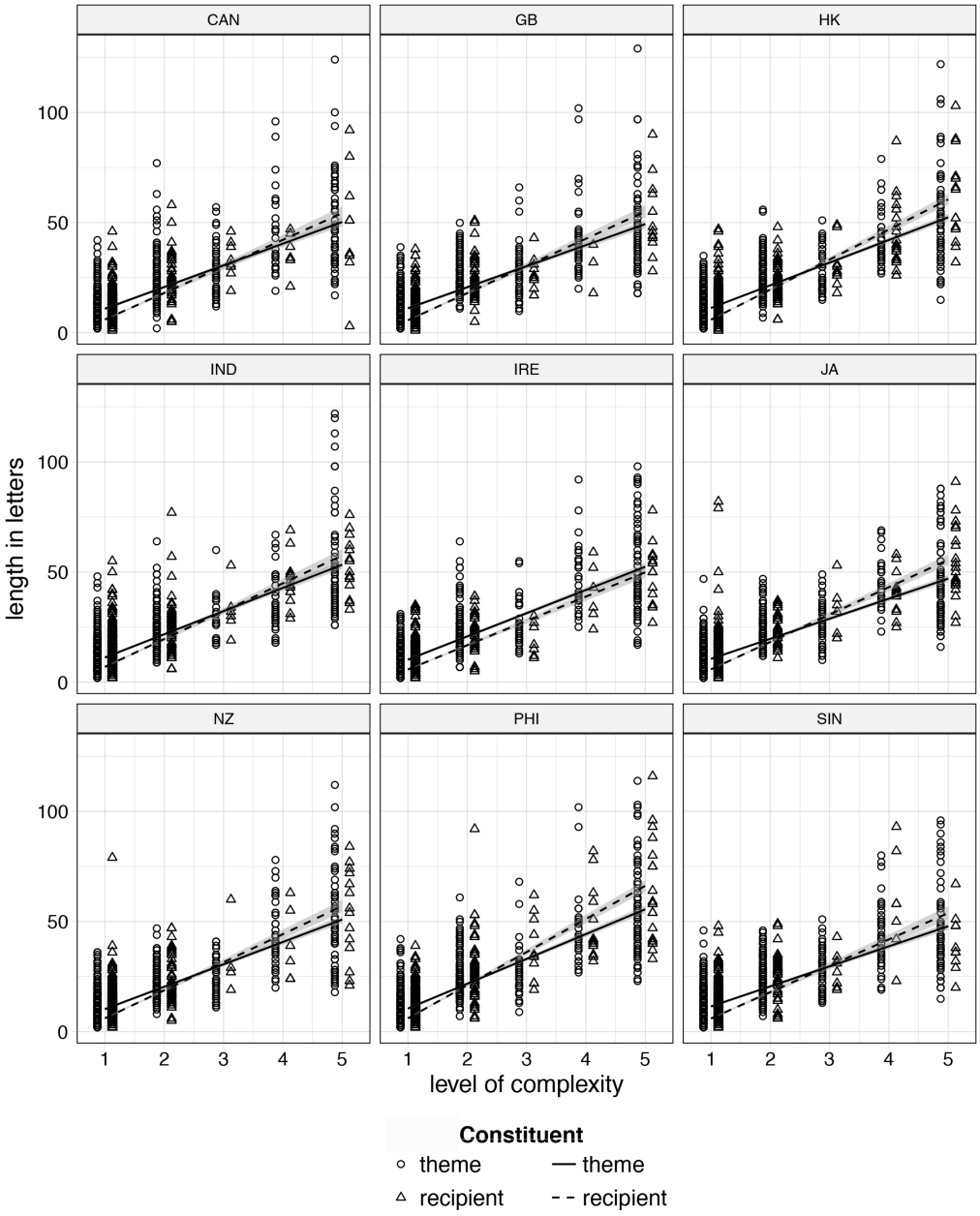


Figure 5.13 Interdependence of structural complexity and length of recipients (dashed line) and themes (solid line) as a smooth regression line — Individual recipients (triangles) and themes (circles) are plotted by complexity level on the x-axis: 1 = simple, 2 = one nominal post-head dependent, 3 = one sentential post-head dependent, 4 = multiple nominal post-head dependents, 5 = multiple (sentential) post-head dependents.

(34) a. *When I played a recording of her speech to* ***the British Sociolinguist Peter Trudgill*** <ICE-NZ:S2B-038>

b. *the goldrushes gave the city* ***a distinctly entrepreneurial character*** <ICE-NZ:W2E-006>

Figure 5.14 plots simple themes and recipients. As noted, recipient outliers are much longer in the number of letters than theme outliers, that is, even though simple recipients are on average shorter than simple themes (as indicated by the median), there are many more recipient outliers that are comparatively longer than themes.

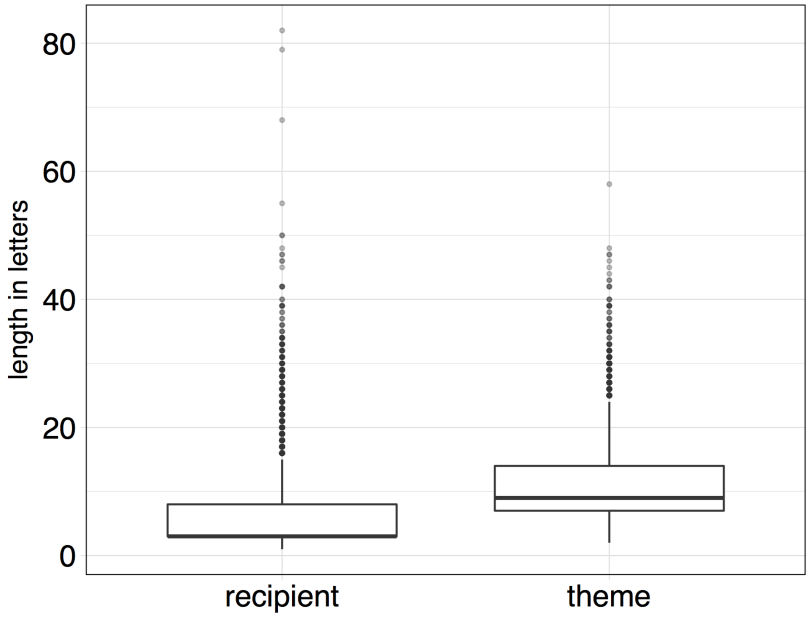


Figure 5.14 Length comparison of simple recipients and simple themes — Simple recipients range much wider in length than simple themes.

Next, in order to gauge the individual impact of NP-structure and NP-length on dative choice, two conditional random forests were fitted to the data with the model formula provided in (35) (see Shih & Grafmiller 2011 for another possibility to tease apart the impact of various weight measurements). Random forests are particularly robust to correlating factors such as length and structural complexity in the data and thus better suited than regression methods to tease apart the importance of each predictor. Five variables were selected at each split and a total of 5000 trees were grown (see also Section 4.4 for a detailed description of the technique). The robustness of the random

forest was confirmed by fitting the same forest with a different random seed. Note that the call structure in (35) includes the predictors WEIGHTRATIO and COMPLEXITYRATIO. The predictor COMPLEXITYRATIO gauges relative complexity of recipient and theme. Relative complexity is calculated by dividing the (numeric) level of the recipient (1-5) by the (numeric) level of the theme (1-5) following the complexity ranking proposed in Table 5.6. If the recipient was simple ('s' = 1) and the theme was followed by one nominal post-head dependent ('spp' = 2), COMPLEXITYRATIO was (1/2=) 0.5. Values of 1 thus indicate that both constituents are equally complex, values below 1 indicate that the theme is more complex than the recipient and values above 1 indicate that the recipient is more complex than the theme.

(35) Variant ~ WEIGHTRATIO + COMPLEXITYRATIO + VARIETY + MODE + VERBSEMANTICS + RECGIVENNESS + THEMEGIVENNESS + RECDEFINITENESS + THEMEDEFINITENESS + RECHEADFREQ + THEMEHEADFREQ + RECTHEMATICITY + THEMETHEMATICITY + PRIMETYPE + RECPRON + THEMEPRON + RECANIMACY + THEMEANIMACY + TYPETOKENRATIO

Variable importance was again calculated using the `varimpAUC()` function from the party package (Janitza et al. 2013) which uses the *C*-statistic instead of predictive accuracy to determine permutated variable importance.

The final forest had an accuracy of 90.2% and a *C*-statistic of 0.96, indicating that the random forest discriminated well between the two dative variants. Variable importance of the forest shows that WEIGHTRATIO is more important than COMPLEXITYRATIO (Figure 5.15) which might be explained by the fact that WEIGHTRATIO accounts for differences in premodifications, something that COMPLEXITYRATIO does not do. The nevertheless high importance of COMPLEXITYRATIO (ranked third in variable importance) highlights that the effect of NP-structure on dative choice is (partially) independent of length.

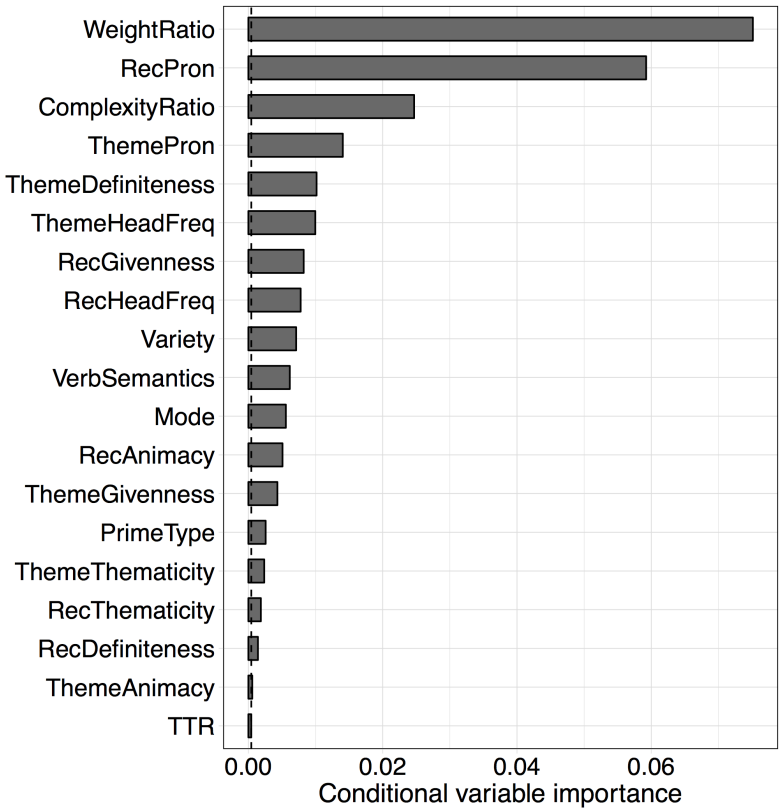


Figure 5.15 Variable importance of the NP-length and NP-structure in a random forest — NP-length (WEIGHTRATIO) is more important than NP-structure (COMPLEXITYRATIO).

5.4.3 The (inexistent) regional malleability of NP-structure

In a next step, the regional variation in the effect of NP-structure was examined by first looking at regional differences in variable importance and second, by assessing the degree of cross-regional malleability of the effect size of NP-structure.

To start with, regional variation in variable importance was gauged by fitting a conditional random forest per variety using the same call structure as in (35) ($mtry = 5$, $ntree = 5000$). Figure 5.16 shows the importance of each predictor by variety. WEIGHTRATIO and recipient pronominality are the two most important predictors across all nine varieties with recipient pronominality being more important than WEIGHTRATIO only in Irish and Indian English. The constraint ranking shown here thereby deviates from the constraint ranking without COMPLEXITYRATIO and with the full dataset (see Section 5.2.2) where recipient pronominality was more important than WEIGHTRATIO only in Indian English. Also note that COMPLEXITYRATIO is the third most important predictor in six varieties, that is, in British, New Zealand,

Jamaican, Singapore, Hong Kong and Philippine English. In Canadian English the third most important predictor is recipient givenness followed by complexity; in Irish English it is recipient head frequency followed by complexity; in Indian English the third most important predictor is recipient head frequency followed by recipient givenness and only then complexity. The by-variety constraint ranking thus shows that NP-structure is not equally important in all varieties.

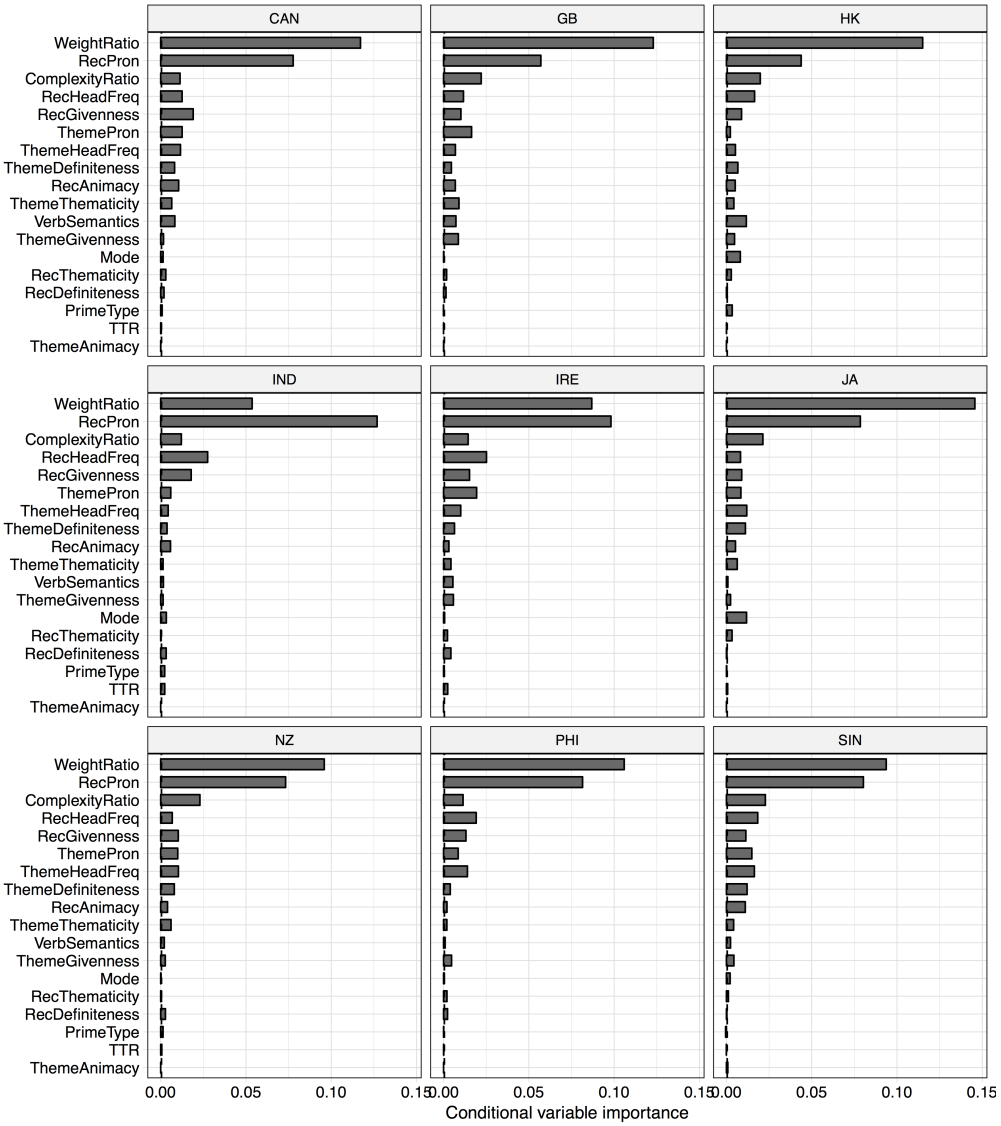


Figure 5.16 Variable importance of NP-length and NP-structure in random forests fitted by variety — NP-structure is often but not always the third most important predictor. Factors are ordered decreasingly by global mean of importance.

To assess the extent to which the effect of COMPLEXITYRATIO is regionally variable, two mixed-effect models were fitted to the data. The relative predictors WEIGHTRATIO and COMPLEXITYRATIO were again used because the levels for recipient and theme complexity were too sparsely distributed across the varieties to make a regression analysis feasible. The models included random intercepts for GENRECOARSE, SPEAKERID, VERB and an interaction of VARIETY with ratio of NP-length (for the first model) or NP-structure (for the second model). Each model made use of contrast coding for VARIETY and did not consider any other predictors. As in previous models, WEIGHTRATIO was log-transformed, standardised and centred around the mean. The call structure is provided in (36). Predictions are for the prepositional dative.

(36) Variant ~ (1|GENRECOARSE) + (1|SPEAKERID) + (1|VERB) + VARIETY * WEIGHTRATIO OR COMPLEXITYRATIO

The performance of the two models was compared regarding each model's *C*-statistic, its accuracy and the variance accounted for by the model. Variance accounted for by the model is expressed with two statistics: Marginal R^2 (R^2_m) indicates variance accounted for by fixed effects as a proportion of the sum of all the variance components, conditional R^2 (R^2_c) also includes the random structure in this calculation. The comparison substantiates the findings from the random forests (see Table 5.7): First, NP-length varies in its effect size across varieties (here: JamE) while NP-structure does not. And second, the model with NP-length performs better than the one with NP-structure: *C*-statistic, accuracy and variance accounted for by the model is higher in the model with NP-length compared to the model with NP-structure (*C*-statistic: 0.967 vs. 0.928; accuracy: 91.17% vs. 85.28%; R^2_m: 0.436 vs. 0.265; R^2_c: 0.830 vs. 0.746). In other words, NP-length is a better predictor of dative choice than NP-structure across all varieties.

Table 5.7 Summary statistics of regression models with NP-structure (COMPLEXITYRATIO) or NP-length (WEIGHTRATIO) in interaction with VARIETY

	WEIGHTRATIO	**COMPLEXITYRATIO**
Cross-varietal deviances	JamE	none
C-statistic	0.967	0.928
Accuracy (in %)	91.17%	85.28%
Variance accounted for	$R^2_m = 0.436$; $R^2_c = 0.830$	$R^2_m = 0.265$; $R^2_c = 0.746$

5.4.4 Nominal constituents only

Since pronominal constituents are by nature short and structurally simple, the next analysis aims to take stock of the impact of the probabilistic constraints on dative choice if the most influential cases are excluded. The extent to which NP-length and NP-structure influence dative choice independently of each other and the extent of their regional malleability was thus explored in a delimited dataset. To that end, I restricted the analyses to observations with only nominal constituents, excluding any dative variants with pronominal themes or recipients ($N = 3{,}099$).

The relative importance of NP-length and NP-structure was again gauged by fitting a random forest on the reduced (nominal) dataset. The random forest included the same predictors as (35) (with the exception of constituent pronominality); hyperparameters were set to *mtry* = 5 and 3000 trees. The robustness of the forest was again confirmed with a different random seed. Variable importance was calculated with the `varimpAUC()` function in the party package (Janitza et al. 2013). As shown in Figure 5.17, excluding pronominal recipients and themes yields a very similar picture to the one in Figure 5.15: WEIGHTRATIO remains more important than COMPLEXITYRATIO as a predictor of dative choice and they both impact the variation partially independent of each other. Furthermore, the exclusion of pronominal recipients and themes leads to a higher rank (and hence greater importance) of the factor VARIETY. In other words, by excluding pronominal constituents, regional differences between varieties gain in importance in the choice of dative variant.

Next, regional malleability of both WEIGHTRATIO and COMPLEXITYRATIO was assessed following the same procedure outlined previously. Two models were fitted, one each with WEIGHTRATIO and COMPLEXITYRATIO. The models included random intercepts for GENRECOARSE, SPEAKERID, VERB and an interaction of VARIETY with ratio of NP-length (for the first model) or NP-structure (for the second model). Each model made use of contrast coding for VARIETY and did not consider any other predictors. As in previous models, WEIGHTRATIO was log-transformed, standardised and centred around the mean. The model formula is repeated here for convenience in (37). Predictions are again for the prepositional dative.

(37) Variant $\sim$ (1|GENRECOARSE) + (1|SPEAKERID) + (1|VERB) + VARIETY $*$ WEIGHTRATIO OR COMPLEXITYRATIO

Summary statistics indicate that both models fit the data well (accuracy for WEIGHTRATIO = 83.7%; accuracy for COMPLEXITYRATIO = 84.2%) and can discriminate

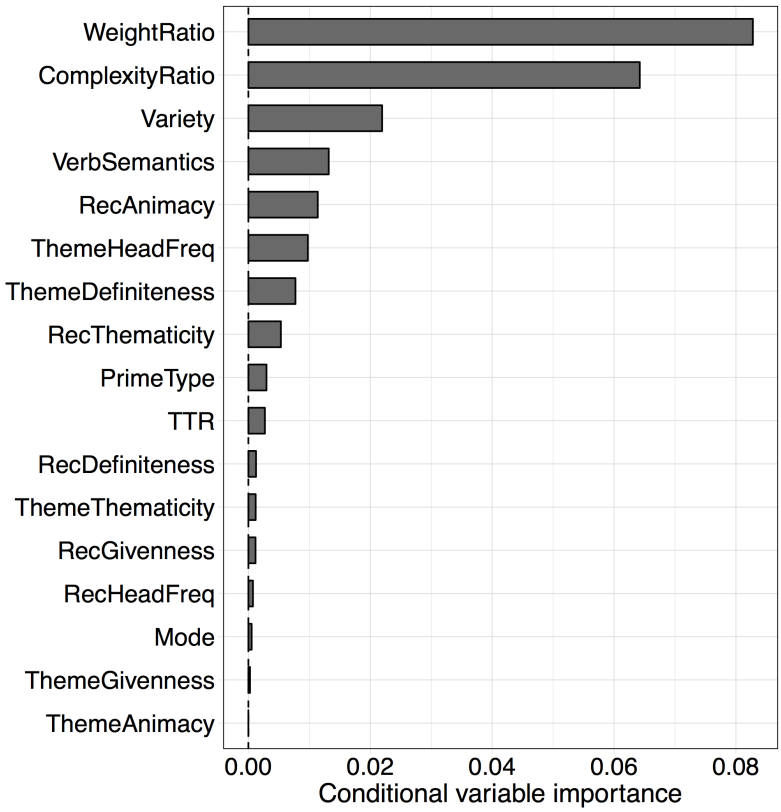


Figure 5.17 Variable importance of the predictors in the random forest analysis with only non-pronominal themes and recipients — NP-length (WEIGHTRATIO) is more important than NP-structure (COMPLEXITYRATIO).

between the variants (*C*-statistic for WEIGHTRATIO = 0.910; *C*-statistic for COMPLEXITYRATIO = 0.916). The results show that WEIGHTRATIO is regionally variable in IrE, JamE and NZE, while COMPLEXITYRATIO is not. In sum, if pronouns are excluded from the analysis, the model with COMPLEXITYRATIO performs slightly better than the one with WEIGHTRATIO. The cross-regional effect of the predictors remains the same as in the full dataset: WEIGHTRATIO is regionally variable, even more so than before, the effect of COMPLEXITYRATIO is, however, still stable across varieties.

5.4.5 Is a five-level predictor necessary?

Since COMPLEXITYRATIO did not turn out to be regionally malleable nor more important than NP-length, the question remains whether a five-level predictor of structural complexity is really necessary or whether the binary predictor used in previous analyses would suffice. This issue was addressed by comparing mixed-effects models and

conditional random forests that included the five-level predictors RECCOMPLEXITY5 and THEMECOMPLEXITY5 with models that included the binary predictors. The binary predictors are renamed for the purpose of this comparison to RECBINCOMPLEXITY and THEMEBINCOMPLEXITY. Recall that the binary predictors only distinguish between 'simple' constituents without any post-head dependents and 'complex' constituents with (an undefined number of) post-head dependents.

The mixed-effect models fitted included random intercepts for GENRECOARSE, SPEAKERID and VERB as well as the fixed effects of RECBINCOMPLEXITY and THEMEBINCOMPLEXITY, and RECCOMPLEXITY5 and THEMECOMPLEXITY5 respectively. Note that this analysis does not make use of the relative predictor COMPLEXITYRATIO. Both models are evaluated based on their AIC (Akaike Information Criterion), their *C*-statistic, accuracy and variance accounted for by the model. Predictions are for the prepositional dative. A model's AIC provides the log-likelihood of model fit and penalises for the number of constraints in the model. AICs thus offer a direct comparison of model fit.

The random forests fitted to the data share the same model formula as previous forests (see 35) but use separate complexity factors instead of relative complexity. The two forests include either the two- or the five-level predictors in their modelling. Parameters are again set to 5,000 trees and $mtry = 5$. Similar to the mixed-effects models, the *C*-statistic and accuracy of each model was calculated to enable a comparison between the two models. The robustness of each random forest was confirmed with a different random seed.

Summary statistics of each of the four models are presented in Table 5.8. The test statistics point out that, overall, a model with a five-level predictor performs better than a model with a two-level predictor for theme and recipient complexity. Zooming in on the mixed-effects models (columns one and three), the model with the five-level predictors has a better (i.e. lower) AIC, higher *C*-statistic, marginally lower accuracy and has more variance accounted for (both conditional and marginal). Moving on to the random forests (columns two and four), the model with the five-level predictors has again a higher *C*-statistic and better accuracy (see Table 5.8). In sum, distinguishing between five instead of only two levels of structural complexity increases the fit of the model. That is, the more fine-grained the weight measure, the better the model. Since the difference between the models is very marginal, I refrained from annotating the dative observations sampled from GloWbE for NP-structure. The comparison between the two mixed-effects models, where NP-structure

constitutes the only fixed effect, especially highlights that even in the absence of any other constraints (for instance, givenness, definiteness, pronominality), the binary predictor of complexity is sufficient to achieve a good model fit.

Table 5.8 Comparison of models with two- and five-level predictors of NP-structure — The comparison supports the use of a five-level predictor of NP-structure.

	two-level predictor		**five-level predictor**	
	GLMER-Model	CRF-Model	GLMER-Model	CRF-Model
AIC	7824.355	n.a.	7840.37	n.a.
C-statistic	0.930	0.957	0.931	0.959
Accuracy	85.64%	89.74%	85.44%	89.91%
Variance	$R^2_m = 0.155$; $R^2_c = 0.703$	n.a.	$R^2_m = 0.185$; $R^2_c = 0.711$	n.a.

The comparison between the four models highlights that the more fine-grained the predictor, the more variance it can account for in the data, which might very well also explain the higher importance of NP-length compared to NP-structure in the previous analyses. As a numeric predictor, NP-length (WEIGHTRATIO) distinguishes among minuscule steps of increasing length in contrast to NP-structure (COMPLEXITYRATIO) which makes quite a coarse distinction between five steps (1-5). What is more, NP-length accounts for differences in the number of premodifying elements while NP-structure does not, offering only one level for premodified constituents without post-head dependents, namely 's' (simple).

While it is beyond the scope of the current study to compare the effects of various measurements of NP-structure and NP-length in more detail here, it remains desirable of future work to address this issue further, for instance by including the even more fine-grained original 16-level predictor of recipient and theme complexity and contrasting it with the effects of NP-length.

5.4.6 Interim summary

The aim of the current section was to expand on Berlage's study by (1) determining the importance of structural complexity and length on dative choice separately and thus to address the question whether length is indeed more important than structural complexity, and (2) to investigate the extent to which syntactic complexity is variable in its effect size across regional varieties of English. In a third step, the analyses

were repeated on a dataset restricted to nominal constituents only in order to control for the highly important effect of pronominal constituents. Following and adapting Berlage's complexity coding to the dative tokens sampled from ICE rendered a scale of complexity that included five levels from least to most complex. Recipient and theme complexity was thereby gauged separately (RECCOMPLEXITY5, THEMECOMPLEXITY5) but also relatively (COMPLEXITYRATIO). The analyses highlight six important findings: First, if both complexity and length are included in a model, length is the more important predictor on a global as well as variety-specific level. Second, the effect of NP-length is cross-regionally malleable while NP-structure is not. Third, the importance of both NP-length and NP-structure is cross-regionally malleable. Fourth, if pronominal constituents are excluded from the analysis, regional differences become more important and fifth, length remains regionally malleable (even more so than before) while complexity is still stable in its effect size. Finally, the more fine-grained five-level predictor of NP-structure leads to a better model fit than a binary predictor of NP-structure that only distinguishes between simple and complex constituents.

The stability of the effect of NP-structure as operationalised in the present study across regionally distinct varieties of English is somehow surprising given the facts that NP-structure is highly correlated with NP-length and that NP-length is indeed regionally variable in its effect size. It might thus be possible that the operationalisation of NP-structure is not sufficient to gauge end-weight effects as well as NP-length does – a proposition that finds support in the more fine-grained nature of NP-length. In any case, the analysis presented here confirms previous findings: Complexity does not turn out to be epiphenomenal to length but plays its own significant role. NP-length is more important than NP-structure in the dative alternation, corroborating findings in Berlage (2014) and Shih & Grafmiller (2011).

5.5 Assessing regional differences in the lexical profiles of the English dative alternation

Besides the regional variability of length effects explored in more detail in the previous section, the mixed-effects model in Section 5.3.4 also pointed to the cross-regional malleability of recipient pronominality. Taking this as a starting point, the present section aims to assess the extent to which differences in lexical instantiations of recipients, themes and verbs can account for the observed regional malleability of probabilistic constraints, specifically with regard to recipient pronominality. To

that end, this section zooms in on the regional variability of the lexical profiles of ditransitive and prepositional datives. Lexical profiles are thereby defined as the system of lexical items (e.g. verbs, recipients, themes) that are mutually attracted within and to a syntactic variant. Measuring the strength of association between lexical items (*collexemes*) and between lexical items and the constructions they occur in (*collostructions*), the present section provides an explanatory link between probabilistic grammars and the structural patterns that speakers are exposed to.

The notion of collostructions draws on some fundamental concepts in Cognitive Linguistics and Construction Grammar. As in Construction Grammar, the current study assumes that lexicon and grammar are not stored separately in a speaker's linguistic knowledge but as intertwined parts of a complex network of hierarchically related *constructemes* – combinations of lexical items and syntactic constructions on various levels of abstractedness. These constructemes can take anywhere from a fully abstract form in which no syntactic slot is lexically predefined (as in the abstract ditransitive constructeme 'Subj V NP NP') to fully instantiated forms where every syntactic slot is lexically filled (as in the concrete ditransitive constructeme *John gives Mary the apple*). Since enough exposure to the same or related constructemes can result in the gradual generalisation of underlying probabilistic constraints (as probabilistic approaches to grammar assume, see, for instance, Bresnan 2007; Szmrecsanyi et al. 2017), measuring the diversity and range of lexical items associated within and with either of the two variants (i.e. the variant's lexical profile) offers an additional window into cross-lectal differences in probabilistic constraints as well as the degree of *collostructional nativisation* in a variety (Mukherjee & Gries 2009). Collostructional nativisation accounts for the emergence of new forms and structures at the level of quantitative shifts in the association between lexical elements and the constructions they occur in and thus offers an indication of the extent to which speakers of different national varieties of English have indigenised their English structurally. Hence, the current section takes stock of the extent to which regional variation in the types of construction that profile the dative variants in English – the diversity and range of lexical items associated within and with either of the two dative variants – can possibly account for the effect of probabilistic indigenisation observed by Bresnan & Hay (2008), Szmrecsanyi et al. (2016), Röthlisberger et al. (2017) and others.

In order to identify a variant's lexical profile, the association strength of covarying collexemes on the syntagmatic level (*collocations*) and of lexicogrammatical covariances on the paradigmatic level (*collostructions*) will be measured using a collexeme

analysis. Mathematically, a collexeme analysis pits observed against expected frequencies of a word co-occurring with another word (in the case of collocations) or with a construction (in the case of collostructions) and uses distributional statistics as outlined in Table 5.9 for collocations (covarying collexeme analysis) and in Table 5.10 for collostructions (distinctive collexeme analysis) to calculate the strength of association between the two linguistic items. Covarying and distinctive collexeme analyses are explained in more detail in the relevant sections.

Table 5.9 Covarying collexeme analysis

	word M in slot 2	all other words in slot 2
word L in slot 1	Freq (L_{slot1} + M_{slot2})	Freq (L_{slot1} + ¬ M_{slot2})
all other words in slot 1	Freq (¬ L_{slot1} + M_{slot2})	Freq (¬ L_{slot1} + ¬ M_{slot2})

Table 5.10 Distinctive collexeme analysis

	construction A	construction B
word L	Freq (L + A)	Freq (L + B)
all other words	Freq (¬ L + A)	Freq (¬ L + B)

The present section is divided into three parts: Section 5.5.1 presents the results of a covarying collexeme analysis that gauges the mutual attraction between the three lexical items within the ditransitive and prepositional dative on the syntagmatic level. Section 5.5.2 assesses the degree to which recipients, themes and verbs are associated with one of the two constructional variants in opposition to the alternating variant. The last section offers a brief summary and some concluding remarks on the lexical profile of the English dative alternation.

5.5.1 Collocations of verb, recipient and theme

In order to gauge the strength of the association between the verb, theme and recipient in either the ditransitive or prepositional dative, I made use of an item-based covarying collexeme analysis (Stefanowitsch & Gries 2005: 9, 23). A covarying collexeme analysis pits the observed frequency of co-occurrence between two lexical elements against the expected frequency given the co-occurrence of other lexical elements within the same construction and measures the strength of association using the base-ten logarithm of the *p*-value from the Fisher-Yates Exact test (Stefanowitsch

& Gries 2005: 9). The resulting log transformed *p*-values can thus range from -∞ (strong repulsion) to 0 (no relation) to +∞ (strong attraction) (Stefanowitsch & Gries 2005: 7). Log transformed *p*-values above 1.30103 are significant at $\alpha = .05$ (since log10(0.05) = -1.30103), and values above 2 and 3 are significant at $\alpha = .01$ and $\alpha = .001$ respectively (Stefanowitsch & Gries 2005: 7). Positive or negative signs of the resulting *p*-values have to be manually adjusted depending on whether the observed frequency is higher (positive association) or lower (negative association) than the expected frequency. Despite objections to using a *p*-value to measure strength of association, Stefanowitsch & Gries (2003) convincingly argue that many of the previously proposed measures involve distributional assumptions about the data which can hardly be met (such as normal distribution and homogeneity of variances) (Stefanowitsch & Gries 2005: 7). Furthermore, some association measures tend to overestimate association strengths of very rare collocations (e.g. the mutual information score), or are unreliable given the fact that most corpus-linguists work with extremely sparse data (Stefanowitsch & Gries 2003: 217-218). The Fisher-Yates Exact test does not have any of these shortcomings – its only disadvantage is the computationally intensive costs to calculate it (Gries & Stefanowitsch 2004: 101).

Table 5.11 presents an example of such covarying collexemes in the ditransitive dative (taken from the dataset of Irish English). In this example, the collocation involves the verb *give* and the theme *it*. The distributional statistic of these two collexemes results in a *p*-value of .542 and a log transformed *p*-value of .266, which is indicative of a rather small and non-significant association between the two lexemes in the ditransitive dative in Irish English.

Table 5.11 The distribution of *give* and *it* in the ditransitive dative in Irish English

	it	other recipients	row totals
give	38	558	596
other verbs	1	327	328
column totals	39	885	924

Following the *Principle of Semantic Coherence* (Stefanowitsch & Gries 2005: 11), the different lexemes that fill the slots in the two dative variants should be semantically compatible not only with the variant they occur in but also with the other lexemes used in that particular construction. A covarying collexeme analysis thus also provides some indication of the meaning associated with a dative variant.

Since a covarying collexeme analysis can only always be computed between two lexical elements (and not more), the three possible combinations of verb, theme and recipient are presented in what follows. For the purpose of this analysis, the lemmas of the verbs, theme heads and recipient heads were used (see Stefanowitsch & Gries 2005: 5).

Verb-recipient collocations

The top figure in Figure 5.18 plots the mean in collocational strength between the verb and recipient in both the ditransitive (dark bars) and the prepositional dative (light bars) across all nine varieties (with confidence intervals). Raw (mean) values are given at the top of each bar. The mutual attraction between verb and recipient is statistically significantly stronger on average in the prepositional dative than in the ditransitive dative in all nine varieties as indicated by an unpaired t-test ($p<.001$ for all nine comparisons after Bonferroni correction). Regarding differences between varieties, mean collocational strength is slightly higher in IndE compared to the global average, a difference that is not statistically significant ($M_{\text{IND}} = 0.716$, $M_{\text{global\_rest}} = 0.583$, $t(254.6) = 1.84$, $p = .6658$).

Turning to collexemes with maximum collocational strength, the pattern becomes a bit more heterogeneous. Table 5.12 lists the three collexemes per variant and variety with the highest collocational strength. Association strength is given in brackets; the top three are shown for illustration purposes.

The top three strongest collexemes in the ditransitive dative often contain a pronominal recipient in contrast to the prepositional datives in which it is foremost a nominal recipient. In cases where a verb strongly collocates with a pronominal recipient in the prepositional dative, the verb itself already has a strong bias towards the prepositional dative (see Section 5.3.2 on the random effects from the mixed-effects model). A closer look at the actual instantiations reveals that most of these strongly collocating verbs and recipients in the prepositional dative often also co-occur with the same theme. For instance, all collocations of *render* and *gospel* (CanE) in the prepositional dative include the theme *obedience* to form the construction *render obedience to the gospel*; all collocations of *charge* and *customer* (BrE) combine to form the construction *charge high prices to customer(s)*. In other words, most of these highly collocational items seem to be quite idiomatic, describe the same event, and/or are used in a specific context.

Table 5.12 Top three covarying verbs and recipients with the strongest association per variant and variety — Association strength is provided in brackets.

Variety	Ditransitive	Prepositional
CanE	*give it* (5.68)	*assign Association* (6.39)
	show you (5.37)	*render gospel* (5.69)
	permit Association (5.19)	*teach student* (5.39)
BrE	*cause defender* (5.91)	*charge customer* (4.09)
	pay VooServers (5.57)	*bring end* (3.92)
	give it (5.39)	*submit employer* (3.69)
HKE	*drop Carolyn* (6.83)	*recommend other* (14.04)
	show you (6.48)	*send me* (8.71)
	pay her (6.23)	*bring Hong Kong* (5.48)
IndE	*promise Louise* (5.14)	*pay company* (6.15)
	give it (3.46)	*sell Pakistan* (4.38)
	wish you (3.39)	*pay Nehru* (4.30)
IrE	*give it* (5.97)	*extend family* (4.30)
	charge section (4.85)	*bring dealer* (3.78)
	show you (4.11)	*submit Minister* (3.60)
JamE	*grant colony* (4.12)	*hand Lodge* (6.42)
	grant Escape (4.12)	*allocate Brown* (4.42)
	drop me (3.90)	*bring community* (3.66)
NZE	*give it* (5.68)	*recommend friend* (7.00)
	show you (5.01)	*show hairdresser* (5.04)
	wish you (4.96)	*demonstrate nation* (4.49)
PhiE	*show student* (8.32)	*deliver government* (5.20)
	send me (4.56)	*forward arm* (5.11)
	bring you (3.63)	*deny that* (4.64)
SinE	*give it* (6.54)	*pass Kalthom* (4.24)
	wish you (4.87)	*assign group* (4.05)
	pay worker (4.68)	*submit Department* (3.45)

The strongest collexemes per variety and variant are exemplified with sentences from the respective varieties in (38) to (46). Examples of the ditransitive dative are given in (a), examples of the prepositional dative in (b).

(38) CanE

a. *I don't even want to* ***give it*** *a mark.* <GloWbE-CAN:B:3336736>

b. . . . ***assign*** *four 4 three-credit course remissions to* ***the Association*** . . . <ICE-CAN:W2D-002>

(39) BrE

a. *Passes to the back of the defence always* ***cause defenders*** *problems.* <ICE-GB:W2D-015>

b. *Shops and garages have been given the go-ahead to* ***charge*** *higher prices to* ***customers****...* <ICE-GB:S2B-019>

(40) HKE

a. ***Drop Carolyn*** *a line ...* <ICE-HK:W2B-032>

b. *I would* ***recommend*** *this book to* ***others****.* <GloWbE-HK:B:3578251>

(41) IndE

a. *... he had* ***promised his wife Louise*** *all happiness ...* <ICE-IND:W1A-018>

b. *why we are* ***paying*** *the full amount to* ***the company*** *...* <GloWbE-IND:G:852887>

(42) IrE

a. *Temper* ***gives it*** *conviction.* <ICE-IRE:W2C-003>

b. *We, the Castlebar Celtic Senior Women's team, wish to* ***extend*** *our deepest sympathies to* ***Jeremy's family*** *...* <GloWbE-IRE:B:3846084>

(43) JamE

a. *... the contingencies of war made it impossible for the British government to* ***grant the colonies*** *unrestricted access to their own reserves ...* <ICE-JA:W2A-006>

b. *I* ***handed*** *it to* ***Detective Inspector Lodge*** *later that day.* <ICE-JA:S1B-063>

(44) NZE

a. *She placed her hand over mine and* ***gave it*** *a squeeze.* <ICE-NZ:W2F-020>

b. *Would* ***recommend*** *Ace to all my* ***friends****.* <GloWbE-NZ:G:1739768>

(45) PhiE

a. ***Show students*** *the worksheet.* <GloWbE-PHI:G:1095064>

b. *... let's* ***deliver*** *a message to* ***the government*** *...* <ICE-PHI:S1B-051>

(46) SinE

a. *But my cousin refused to eat it after* ***giving it*** *a single lick.* <GloWbE-SIN:G:1033746>

b. *Just* ***pass*** *the 2 boxes to* ***Kalthom****.* <ICE-SIN:W1B-011>

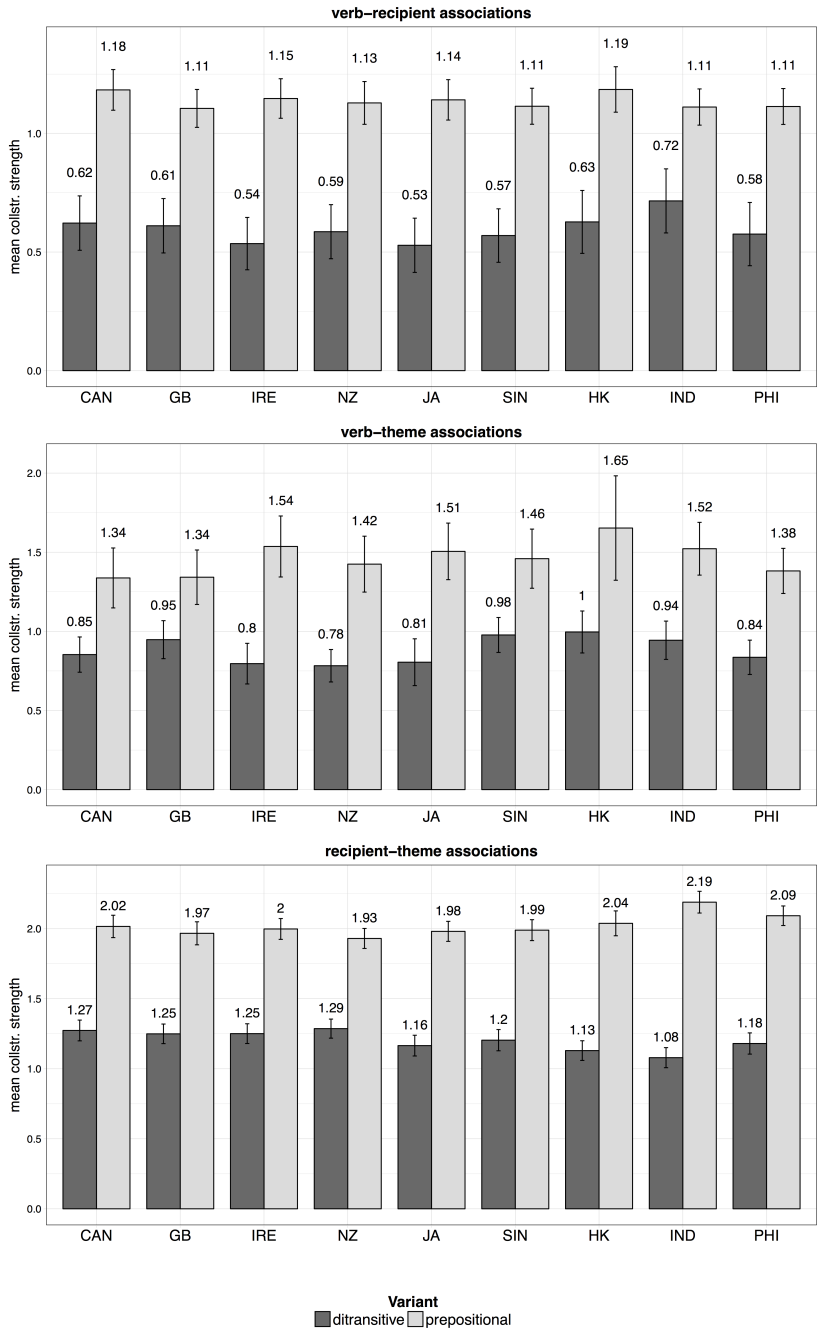


Figure 5.18 (a) top: Mean collocational strength between verb and recipient by variant; **(b) middle:** Mean collocational strength between verb and theme by variant; **(c) bottom:** Mean collocational strength between recipient and theme by variant — All graphs include confidence intervals. Mean strength is statistically significantly higher in the prepositional dative than in the ditransitive dative. No statistically significant regional differences emerge. Native varieties appear on the left side, non-native varieties appear on the right side of the graph.

Verb-theme collocations

Next, the association strength between verbs and themes was measured and the mean collocational strength was compared by variant across the nine varieties (see middle figure in Figure 5.18). Again, collexemes are on average more strongly attracted to each other in the prepositional dative (light bars) than in the ditransitive (dark bars). This difference in mutual attraction is statistically significant in all nine varieties as indicated by an unpaired *t*-test ($p<.01$ in BrE and HKE; $p<.001$ in all other varieties after Bonferroni correction). No statistically significant regional differences emerge.

The three strongest verb-theme collocations per variant and variety on a global level mainly express an act of communication (e.g. *tell story, teach lesson, drop line, wish success*) and covary more strongly in the ditransitive dative than in the prepositional dative (with a few exceptions) as indicated by the association strengths provided in brackets in Table 5.13. Some covarying collexemes are strongly associated in both the ditransitive and the prepositional dative (e.g. *write letter, tell story*). Similar association patterns between verbs of communication and the ditransitive dative were also found in Stefanowitsch and Gries (2003: 240 fn. 13). At the same time, the verb *pay* and the theme *attention* are strongly associated with each other in the prepositional dative.

A comparison with the verb-recipient collocations described in the previous section reveals four differences: First, we can observe that the association strength is a bit higher for the top three verb-theme collocations than for the top three verb-recipient collocations. Second, individual verb-theme collocations are used in combination with a large number of recipient types as well as recipient tokens compared to verb-recipient collocations that often co-occur with one or two themes and only contain a handful of tokens. Recall that the top three verb-recipient collocations often co-occur with the same theme. Third, verb-theme collocations are very similar across all nine varieties. And fourth, *pay attention* has the highest collocational strength in the prepositional dative in almost all varieties and hence seems to be some kind of universal prototypical verb-theme collocation for that variant. With regard to similarities between verb-recipient and verb-theme collocations, some verbs associated strongly with recipients in the ditransitive dative are also strongly associated with themes in the ditransitive dative, and verbs associated strongly with recipients in the prepositional dative are also strongly associated with themes in the prepositional dative. These findings substantiate results of earlier works that observed a strong verb bias towards either

Table 5.13 Top three covarying verbs and themes with the strongest association per variant and variety — Association strength is provided in brackets.

Variety	Ditransitive	Prepositional
CanE	*tell this* (15.28)	*pay attention* (18.70)
	wish best (8.02)	*render obedience* (12.67)
	tell story (6.38)	*write letter* (7.36)
BrE	*tell story* (16.06)	*write letter* (15.34)
	tell that (11.22)	*pay attention* (8.25)
	tell this (9.28)	*send message* (7.01)
HKE	*send mail* (18.76)	*pay attention* (55.41)
	bring replay (16.83)	*recommend book* (21.63)
	tell story (14.17)	*send mail* (19.55)
IndE	*drop line* (11.16)	*pay attention* (18.27)
	teach lesson (10.70)	*write letter* (12.04)
	wish success (9.81)	*submit memorandum* (9.34)
IrE	*tell this* (18.61)	*cause damage* (15.30)
	tell that (10.32)	*pay tribute* (9.80)
	send card (10.13)	*bid farewell* (6.84)
JamE	*tell something* (25.23)	*pay attention* (15.43)
	drop line (14.23)	*pay tribute* (9.47)
	wish best (11.17)	*write letter* (8.27)
NZE	*write letter* (14.81)	*pay attention* (17.86)
	wish Christmas (8.02)	*write letter* (13.51)
	send copy (6.55)	*tell story* (6.12)
PhiE	*write letter* (11.39)	*pay attention* (15.93)
	teach lesson (9.99)	*tell story* (8.75)
	tell more (9.62)	*write letter* (6.65)
SinE	*tell story* (13.26)	*pay attention* (22.37)
	tell thing (10.88)	*write letter* (14.80)
	keep company (10.77)	*pass it* (5.59)

of the two dative variants, further corroborating the importance of verbs in dative choice vis-à-vis the influence of theme and recipient.

The strongest verb-theme collocations per variety and variant are, again, exemplified with sentences from the respective varieties in (47) to (55). Ditransitive datives are exemplified in (a), prepositional datives in (b).

(47) CanE
a. *I hate to **tell** you **this** ...* <ICE-CAN:S1A-075>
b. *On boring days, I **pay cursory attention** to world events ...* <ICE-CAN:W2F-019>

(48) BrE
a. *I'll **tell** you **a funny story about working class** later.* <ICE-GB:S1A-037>
b. *At last I put pen to paper and actually **write a letter** to you.* <ICE-GB:W1B-002>

(49) HKE
a. *So do **send** me **mail** before Wed so that I can reply you.* <ICE-HK:W1B-011>
b. *They **pay much attention** to sanitation, ...* <GloWbE-HK:G:1126078>

(50) IndE
a. *Please **drop** me **a line**.* <ICE-IND:W1B-006>
b. *But then he doesn't **pay attention** to all this na all these things ...* <ICE-IND:S1A-037>

(51) IrE
a. *I **tell** residents **this** but the work never happens.* <ICE-IRE:W2C-007>
b. *... it didn't uh **cause damage** to anyone who came on the land ...* <ICE-IRE:S2A-063>

(52) JamE
a. *Let me **tell** you **something**.* <ICE-JA:S1B-046>
b. *Nobody was **paying attention** to Auntie Maggie's size.* <ICE-JA:W2F-018>

(53) NZE
a. *but so we have to **write** her **a letter** before then.* <ICE-NZ:S1A-017>
b. *A few employers **paid little attention** to the trial period.* <GloWbE-NZ:G:1749043>

(54) PhiE
a. *I remember graduation from college he **wrote** me **a letter**.* <ICE-PHI:S1A-005>
b. *Please **pay particular attention** to no. 6 in the guidelines.* <ICE-PHI:W1B-028>

(55) SinE
a. ***Tell** your child **stories about himself** ...* <ICE-SIN:W2D-020>
b. *Then I will also try to **pay more attention** to all of the women.* <GloWbE-SIN:B:3528828>

Recipient-theme collocations

Finally, association strength between recipient and theme heads as well as their mean collocational strength were calculated and compared across the nine varieties (see bottom figure in Figure 5.18). Again, collexemes are on average more strongly

attracted to each other in the prepositional dative (light bars) than in the ditransitive (dark bars). This difference in mutual attraction is statistically significant in all nine varieties as indicated by an unpaired *t*-test ($p<.001$ after Bonferroni correction).

The top three most strongly associated covarying themes and recipients are given in Table 5.14. The picture that emerges is similar to covarying verbs and recipients in that the covarying collexemes with the highest collocational strength count only very few instances. Also, most covarying collexemes seem to form part of a fixed expression and are used in a specific context or text style. Often, these covarying recipients and themes come from the same text. In IndE, HKE and BrE, the top-ranked collocations are found in the prepositional dative in comparison to all other varieties, where the top-ranked collocations are found in the ditransitive dative. A comparison with previous covarying collexeme analyses in this study reveals that the top-ranked covarying recipients and themes and the top-ranked covarying verbs and recipients are often part of the same dative observation.

The strongest recipient-theme collocations per variety and variant are exemplified with sentences from the respective varieties and variants in (56) to (64). Again, ditransitive datives are exemplified in (a) and prepositional datives in (b).

(56) CanE

a. . . . *trust the admissions committee will give* ***her application serious consideration****.* <ICE-CAN:W1B-030>

b. . . . *the Employer agrees to assign* ***five 5 three-credit course remissions*** *to* ***the Association*** *each term.* <ICE-CAN:W2D-002>

(57) BrE

a. *Passes to the back of the defence always cause* ***defenders problems****.* <ICE-GB:W2D-015>

b. *She was just showing* ***it*** *to* ***me****.* <ICE-GB:S1A-047>

(58) HKE

a. *Did it occur to you to pay* ***her five hundred dollars*** . . . <ICE-HK:S1B-066>

b. *I would recommend* ***this book*** *to* ***others****, because it's different from the usual.* <GloWbE-HK:B:3578251>

(59) IndE

a. *On one side he had promised* ***his wife Louise all happiness*** . . . <ICE-IND:W1A-018>

b. . . . *but we are paying* ***the full amount*** *to* ***the company*** . . . <GloWbE-IND:G:852887>

Table 5.14 Top three covarying recipients and themes with the strongest association per variant and variety — Association strength is provided in brackets.

Variety	Ditransitive	Prepositional
CanE	*application consideration* (9.86)	*Association remissions* (6.99)
	Association use (5.67)	*gospel obedience* (5.69)
	dog trick (5.19)	*Reif knife* (4.88)
BrE	*defender problem* (4.68)	*me it* (7.48)
	VooServers fee (4.38)	*you it* (5.15)
	player credit (4.08)	*Rome tax* (4.87)
HKE	*her dollar* (8.77)	*other book* (15.47)
	Carolyn line (6.63)	*me mail* (7.83)
	you example (5.31)	*Director consent* (6.43)
IndE	*Louise happiness* (5.62)	*company amount* (13.47)
	you example (4.76)	*Nehru homage* (7.99)
	baby food (4.62)	*Pakistan aircraft* (7.03)
IrE	*it try* (6.98)	*node colour* (4.78)
	section rent (5.63)	*Major paper* (4.78)
	Towers workout (5.63)	*talk direction* (4.78)
JamE	*himself room* (7.19)	*police name* (6.00)
	Peter message (6.55)	*police statement* (5.27)
	Court jurisdiction (4.90)	*Brown acre* (4.90)
NZE	*God authority* (8.54)	*owner price* (4.49)
	God permission (6.85)	*hairdresser look* (4.49)
	it go (5.35)	*medium information* (3.96)
PhiE	*u room* (6.91)	*principle life* (5.80)
	him ultimatum (6.89)	*me it* (5.29)
	u strength (5.91)	*Cruz concept* (5.11)
SinE	*Sato training* (13.11)	*winner medal* (5.00)
	authority power (6.59)	*Groupon friend* (5.00)
	it try (6.23)	*desire reign* (4.53)

(60) IrE

a. *. . . they were prepared to give* ***it a try****.* <ICE-IRE:S2A-044>

b. *Assign* ***colour I*** *to* ***the first node on the list****.* <ICE-IRE:S2A-037>

(61) JamE

a. *Darryl Brown looking to give* ***himself a bit of room*** *. . .* <ICE-JA:S2A-003>

b. *Several days after that you gave* ***the name*** *to* ***the police****.* <ICE-JA:S1B-069>

(62) NZE

a. . . . *His proclamation that God would send judgment on the evil of the earth gave* ***God authority on earth****.* <GloWbE-NZ:G:1723518>

b. *it says that he paid* ***the price*** *to* ***the owner****.* <ICE-NZ:S1B-017>

(63) PhiE

a. *Come with me and I will show* ***u your room****.* <GloWbE-PHI:G:1119990>

b. . . . *the realm that has enabled us to give* ***life*** *to* ***the principle of a free market place of ideas*** . . . <GloWbE-PHI:B:3558148>

(64) SinE

a. *Just give Sato* ***a little sword training and a costume****,* . . . <GloWbE-SIN:G:1049128>

b. . . . *we're wondering who is gonna present* ***the medals*** *to* ***the winners****.* <ICE-SIN:S2A-018>

In sum, the covarying collexeme analysis shows that the mutual attraction between recipients and themes, verbs and themes and verbs and recipients is highest in the prepositional dative on average across all nine varieties of English. In other words, verbs, themes and recipients used in the prepositional dative are frequently re-used in combination with the same constituents. In ditransitive datives, on the other hand, the lexical constituents seem to be more freely associated with each other on average. At the same time, the most strongly associated covarying collexemes in the ditransitive dative outrank the strongest covarying collexemes in the prepositional dative in association strength. Cross-lectal differences emerge in that regard: The strongest covarying collexemes in IrE always occur in the ditransitive dative (for all three combinations); in HKE and IndE, they always occur in the prepositional dative. Other varieties pattern less consistently.

On a global scale, covarying collexemes with verbs and recipients often also co-occur strongly with the same themes. For instance, CanE includes the same two ditransitive variants in the top three of both categories of verb-recipient collocations and recipient-theme collocations, given in (65) and (66) below (verb, theme and recipient head are highlighted in bold face).

(65) *The Employer shall* ***permit*** *the* ***Association use*** *of suitable meeting rooms free of charge.* <ICE-CAN:W2D-002>

(66) *You can* ***teach*** *old* ***dogs*** *new* ***tricks****.* <ICE-CAN:W2D-014>

The same overlap can be found in the top-ranked collocations in the prepositional dative. In CanE, for example, you find *assign remissions to Association* as well as *render obedience to the gospel*. The few instances in the data of each of these collocations highlight the lexically and often also stylistically fixed character of the construction. Interestingly enough, no such pattern was observed for verb-theme collocations suggesting that these collocations pattern more independently from the recipient. The importance of verb-theme collocations in contrast to collocations involving a recipient finds support from the mixed-effects model fitted in Section 5.3.2. As shown there, the random effects of verb and theme head accounted for a larger portion of variance than the recipient. The analysis has furthermore highlighted that covarying verbs and themes often express an act of communication, confirming results of previous studies (see Stefanowitsch & Gries 2003)

5.5.2 Collostructional associations of verb, recipient and theme

Collostructional or distinctive collexeme analysis is concerned with significant associations between lexical items and their argument roles in a particular construction A compared to a functionally or semantically similar construction B (Gries & Stefanowitsch 2004: 101; Stefanowitsch & Gries 2005: 8). A distinctive collexeme analysis makes use of the same distributional statistic as a covarying collexeme analysis. The analysis pits observed versus expected frequency of lexical items occurring in one variant compared to the other variant and measures the strength of association using the base-ten logarithm of the *p*-value from the Fisher-Yates Exact test (Mukherjee & Gries 2009: 39). A distinctive collexeme analysis can thus give an indication of the extent to which lexical items are more important (stronger association of lexemes to variant) in one variety and variant compared to another variety and variant. For instance, the association of the recipient *me* is much stronger with the ditransitive than with the prepositional dative in Indian English (see Table 5.15). The test statistic for this collostruction results in a log transformed *p*-value of 18.18 which compares to a *p*-value of 6.579456e-19 ($= 10^{-18.18}$) for the association with the ditransitive variant. The relation between the recipient *me* and the ditransitive dative is thus statistically significant at the $p<.001$ level in Indian English. Association strength was calculated in what follows with the `coll.analysis 3.5()` R-script by Gries (Gries 2014).

In the present study, the distinctive collexeme analysis reflects on the lexical diversity associated with one variant compared to the other variant and can further

Table 5.15 The distribution of *me* in the ditransitive dative vs. prepositional dative in Indian English

	ditransitive	prepositional	row totals
me	139	16	155
other recipients	773	622	1395
column totals	912	638	1550

provide an indication of the extent to which this lexical diversity differs from variety to variety. Note also that a distinctive collexeme analysis can only measure the strength of association between a lexical item and the syntactic variant it occurs in compared to this item's repulsion from another predefined (and semantically similar) construction (see Stefanowitsch & Gries 2009: Ch. 5 for a detailed overview). Hence, a strong association between a lexical item and the ditransitive dative is always only indicative of a collostruction in juxtaposition to the prepositional dative and not to the language or variety as a whole.

Collostructional associations of the verb

A vast body of research has shown that the choice of dative verb exerts an important influence on syntactic variation in the dative alternation (e.g. Rappaport Hovav & Levin 2008). Since most previous corpus-based research on the dative alternation primarily focuses on the verb *give* or restricts attention to one or two varieties of English, the number of large-scale comparative studies that investigate the effect of verb choice have so far remained rather sparse (see, for instance, Bresnan & Hay 2008 on NZE and American English; Bresnan & Ford 2010 on Australian and American English). As the current dataset includes a total of 86 alternating verbs across nine varieties of English, more generalisable conclusions can be drawn with regard to the verbs' lexical effect on dative choice.

Results of the distinctive collexeme analysis indicate that the verb is on average more strongly associated with the ditransitive than the prepositional dative across all varieties (see Figure 5.19). In IrE and HKE, the difference in mean collostructional strength between the ditransitive and the prepositional dative is more pronounced than in the other varieties. No significance tests were computed, however, due to overlapping and very wide confidence intervals which are indicative of sparse data (Figure 5.19)

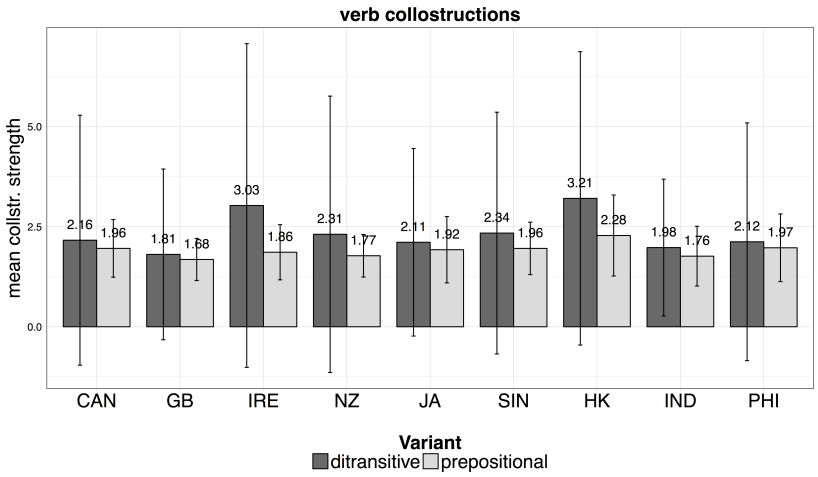


Figure 5.19 Mean collostructional strength of verbs in the ditransitive and the prepositional dative in nine varieties of English (with confidence intervals) — The verb is on average more closely associated with the ditransitive dative than the prepositional dative.

Summarising the three verbs most strongly associated with the ditransitive dative in each variety adds up to a total of six verbs with the highest collostructional strength: *give, teach, tell, wish, show* and *allow*. Among these, *give* is the verb most strongly associated with the ditransitive dative by far in all nine varieties (see top figure in Figure 5.20). Only minor cross-regional differences can be observed: In all but one variety (IndE), the next most strongly associated verb in the ditransitive is *tell*. And association strength of *give* with the ditransitive is highest in SinE, HKE, CanE and NZE.

Association strengths of verbs highly attracted to the prepositional dative are more heterogeneously distributed cross-regionally. All in all, *pay* is the verb most strongly associated with the prepositional dative in six out of nine varieties (see bottom figure in Figure 5.20). In both CanE and BrE, *bring* and *explain* follow *pay* in collostructional strength. IrE and JamE are similar regarding verb type but differ in the verbs' collostructional strength. IndE and PhiE stand somewhat apart from the others as *submit* ranks very high in IndE and is the most strongly associated verb with the prepositional dative in PhiE, in contrast to all other varieties in which *submit* does not constitute a verb strongly associated with the prepositional dative in the top three.

Apart from these minor lexical differences, the varieties further diverge from each other in the prepositional dative with regard to the dispersion of collostruc-

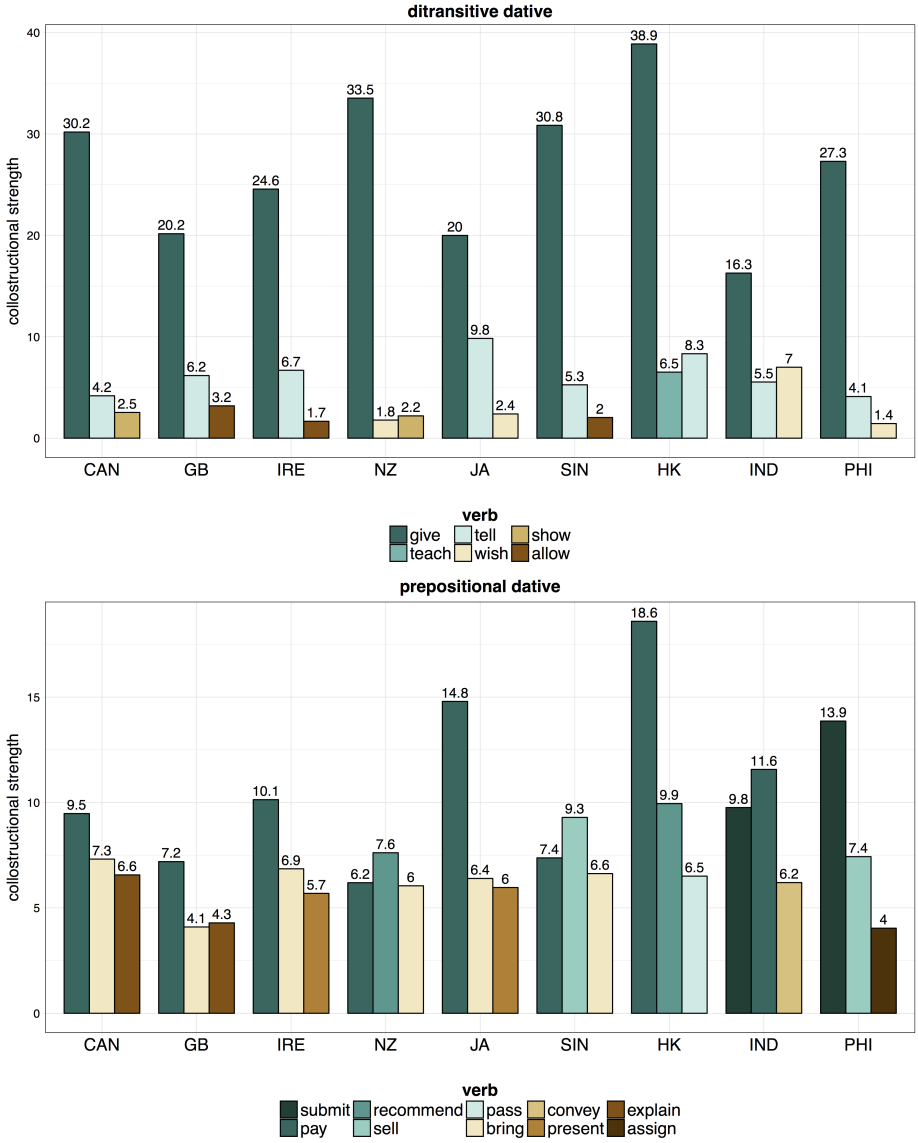


Figure 5.20 (a) top: Collostructional strength of the three verbs most strongly associated with the ditransitive dative — The prototypical ditransitive verb is *give* and association strength is highest in SinE, HKE, CanE and NZE. **(b) bottom**: Collostructional strength of the three verbs most strongly associated with the prepositional dative — Collostructional strength is regionally variably distributed across different verbs. Native varieties appear on the left, non-native varieties on the right side of the graph.

tional strength. Native varieties (the four varieties to the left) exhibit largely similar collostructional strength across the top three verbs (with SinE patterning close to the native varieties). In the non-native varieties, the difference in collostructional

strength between the verbs is much more pronounced with one verb (or two) being very strongly associated with the prepositional dative.

Collostructional associations of the recipient

A collostructional analysis of the recipients does not only offer insights into the lexical instantiations of recipients associated with the ditransitive and prepositional dative but might eventually also provide an explanation regarding the cross-regional variability of the effect of recipient pronominality observed previously.

Mean collostructional strength of all recipients are shown in Figure 5.21 (with confidence intervals). Two differences between native (the four varieties to the left) and non-native varieties (the five varieties to the right) should be noted. First, in native varieties the recipient is on average more closely associated with the prepositional dative than with the ditransitive dative while the association in non-native varieties is stronger with the ditransitive than with the prepositional dative. Second, the average strength of association of recipients and the ditransitive dative is generally higher in non-native varieties compared to native varieties. Association strength with the prepositional dative is marginally lower in non-native varieties compared to native varieties. The wide confidence intervals with ditransitive datives are indicative of more variance in the ditransitive than in the prepositional datives, similar to verb associations.

The higher mean collostructional strength in the ditransitive dative in non-native varieties can be accounted for if we take a closer look at the recipients themselves. The seven recipients most strongly associated with the ditransitive are pronominal recipients in all varieties (see Figure 5.22), namely *you, me, us, them, him, it* and *her* (in decreasing order of collostructional strength). The comparison reveals remarkable regional differences: In non-native varieties, collostructional strength is generally higher than in native varieties. This is most evident in the collostructional strength of *you*, suggesting that the ditransitive dative with *you* is a highly entrenched collostruction in non-native varieties. Speakers of native varieties on the other hand seem to be more flexible with regard to the lexical items that fulfil the role of recipient.

Collostructional strengths of recipients in the prepositional dative are not in any way similarly regionally variable as in the ditransitive dative (not visualised here). Also, the highly disparate mix of recipients associated with the prepositional dative makes a clear visualisation impossible. Instead, the three recipients with the highest

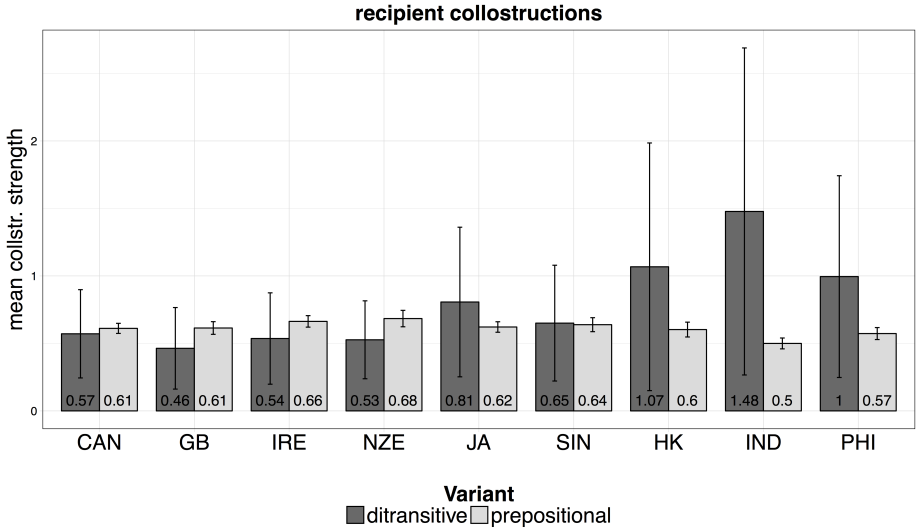


Figure 5.21 Mean collostructional strength of recipients by variant and variety — In native varieties (CanE, BrE, IrE and NZE), the recipient is less attracted to the ditransitive dative than to the prepositional dative. In non-native varieties (JamE, SinE, HKE, IndE and PhiE), the reverse is the case. Native varieties appear on the left, non-native varieties on the right side of the graph.

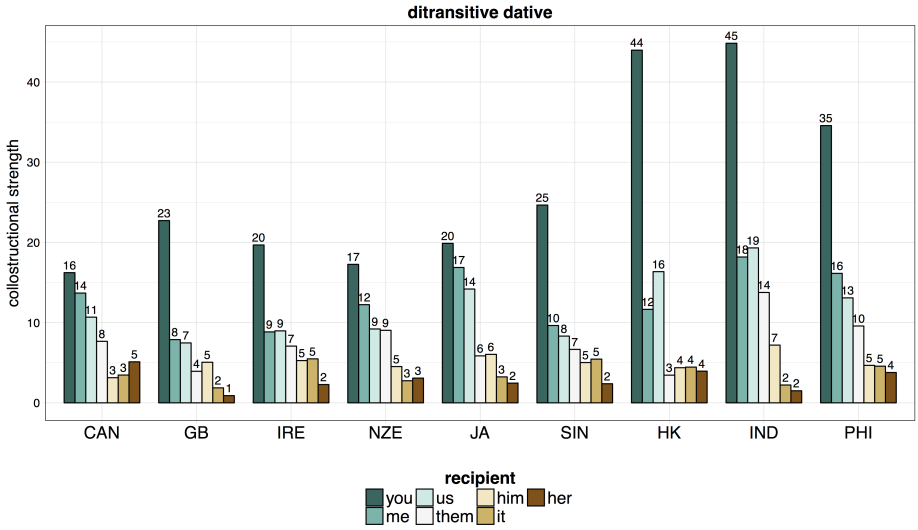


Figure 5.22 Collostructional strength of the top seven recipients in the ditransitive dative per variety — Collostructional strength is generally higher in non-native varieties compared to native varieties. Native varieties appear on the left, non-native varieties on the right side of the graph.

collostructional strength per variety in the prepositional dative are listed in Table 5.16, with collostructional strength in brackets. As shown, recipients strongly associated with the prepositional dative often entail an animate entity (humans or animals) or the demonstrative pronoun *that*.

Table 5.16 Top three recipients most strongly associated with the prepositional dative by variety — Collostructional strength is given in brackets.

Variety	**Recipient**
CanE	*that* (3.27)
	government (1.79)
	people (1.64)
BrE	*that* (4.29)
	people (2.67)
	family (2.14)
HKE	*other* (6.23)
	family (3.10)
	staff (3.10)
IndE	*people* (5.17)
	company (3.10)
	government (2.71)
IrE	*anyone* (2.83)
	that (2.17)
	another (1.70)
JamE	*that* (2.70)
	government (2.16)
	customer (2.04)
NZE	*that* (5.43)
	friend (3.25)
	people (2.20)
PhiE	*people* (4.30)
	company (3.78)
	member (2.13)
SinE	*person* (4.31)
	company (2.99)
	audience (2.69)

Collostructional associations of the theme

In contrast to the lexical effect of the recipient – especially the recipient's association with the ditransitive dative – the lexical effect of the theme is fairly stable across both native and non-native varieties. Mean collostructional strength does not exhibit any major regional differences based on variety type, neither in the ditransitive nor in the prepositional dative (see Figure 5.23). The theme is overall more strongly associated with the prepositional dative in all nine varieties and less so with the ditransitive dative. This difference in association strength between the prepositional and the ditransitive dative is statistically significant as indicated by an unpaired t-test ($p<.05$ for all varieties apart from IrE at $p<.001$, and JamE, NZE and PhiE at $p<.01$). No major splits between native and non-native varieties can be discerned with regard to the theme apart from a slightly stronger association of themes with ditransitive datives in non-native varieties and one significant regional outlier: Collostructional strength in the ditransitive dative is higher in IndE compared to the other varieties – this difference is statistically significant ($M_{\text{IND}} = 0.394$, $M_{\text{global\_rest}} = 0.284$, $t(453.78) = 5.05$, $p<.001$).

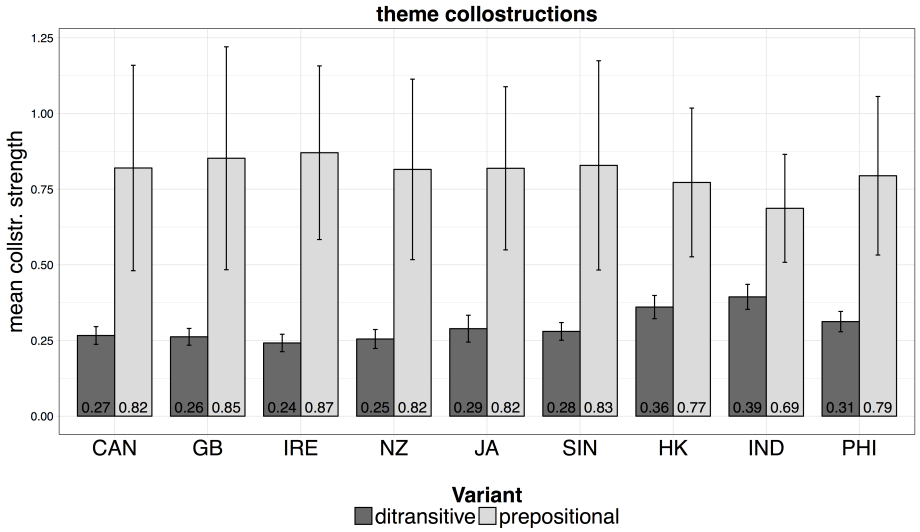


Figure 5.23 Mean collostructional strength of themes by variant and variety — On average, themes are slightly stronger associated with the ditransitive dative in non-native varieties compared to native varieties. Native varieties appear on the left side, non-native varieties appear on the right side of the graph.

Zooming in on the different themes associated with ditransitive datives reveals striking diversity. The ten most strongly associated themes with the ditransitive dative in each variety sum up to a total of 38 different lexical items. Frequently occurring themes

include *chance* and *opportunity* (in the top ten of eight varieties) and *time, example* and *idea* (in the top ten of six varieties). The three most strongly associated themes in the ditransitive in each variety result in nine lexical items, visualised in the top figure in Figure 5.24. The collostructional strength of these nine lexical items is more or less evenly distributed across all varieties.

Similar lexical diversity can be found in the themes that strongly associate with the prepositional dative. Cross-regionally, *it* is the theme most strongly associated with the prepositional dative (see bottom figure in Figure 5.24). Again, comparison of the three themes with the highest association strength in the prepositional dative in each variety does not reveal any cross-regional differences. Apart from Hong Kong English, where *attention* competes with *it* in terms of association strength (see also SinE and IndE where a slightly similar pattern emerges), cross-regional stability prevails.

5.5.3 Interim summary

The present section took regional variability in the effect size of recipient pronominality as a starting point to investigate the extent to which the lexical profiles of the two dative variants are cross-regional malleable. To that end, covarying and distinctive collexeme analyses gauged the strength of association between the lexical items (verb, recipient, theme) within the two variants and the strength of association between the lexical items and the variant in which they occurred.

Measuring the mutual attraction of recipients and themes, verbs and themes and verbs and recipients in either of the two variants revealed that mutual attraction between the constituents is on average higher in the prepositional dative than the ditransitive dative. Strong verb-theme collocations in the ditransitive dative often express acts of communication. Covarying collexemes with the highest collocational strength are found in the prepositional dative primarily in HKE and IndE. What is more, collocations that include a recipient (verb-recipient and theme-recipient collocations) often form part of a seemingly fixed or idiomatic expression.

The distinctive collexeme analysis further revealed that verbs are more closely associated with the ditransitive dative than the prepositional dative on average with very pronounced differences in association strength between the two variants in HKE and IrE. The prototypical ditransitive verb is *give*, especially in SinE, HKE, PhiE and CanE, and *pay* appears to be the prototypical verb of the prepositional dative. Some marginal regional differences can be found with regard to the difference in mean

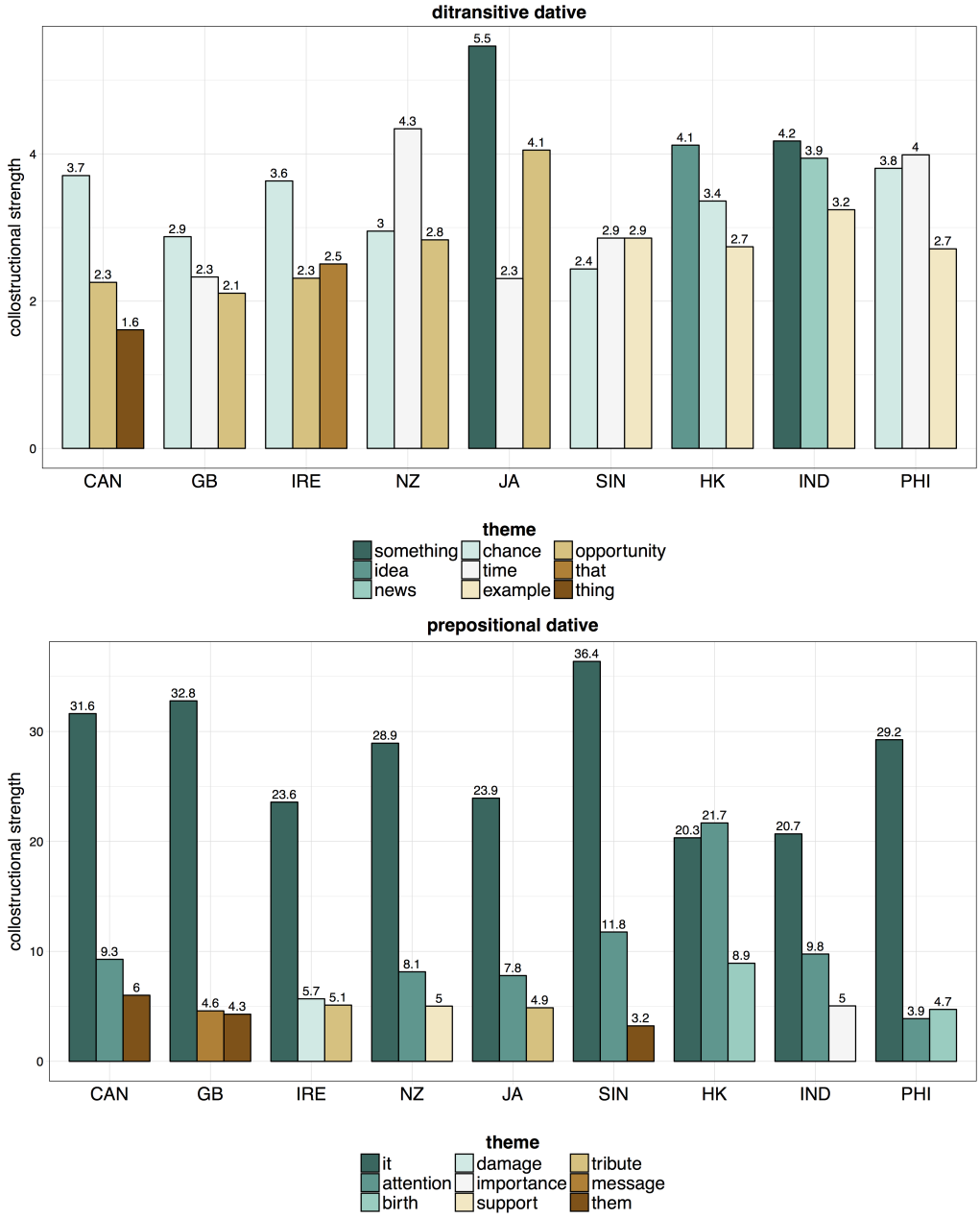


Figure 5.24 (a) top: Top three most strongly associated themes in the ditransitive dative across nine varieties of English — Collostructional strength does not differ extensively cross-regionally. **(b) bottom**: Themes most strongly associated with the prepositional dative across nine varieties of English — Collostructional strength does not differ extensively cross-regionally. Native varieties appear on the left, non-native varieties on the right side of the graphs.

collostructional strength between the variants but regional stability largely prevails. Regional stability also prevails in collostructions with the theme. Mean collostructional strength is overall higher in the prepositional than in the ditransitive dative. *It* is thereby the theme most strongly associated with the prepositional dative. In contrast to the lexical diversity in theme collostructions, recipients are lexically much more restricted. The list of recipients most strongly associated with the ditransitive dative is headed by seven pronominal recipients. Not only is the association between recipients and the ditransitive stronger in non-native varieties compared to native varieties, recipients in native varieties are also more closely associated with the prepositional than with the ditransitive dative.

Speakers of non-native varieties thus seem to follow the principle of recycling lexical items in the ditransitive dative. They associate both the recipient and the theme more closely with the ditransitive dative than speakers of native varieties. This finding finds further support when type-token ratios of recipients and themes in the ditransitive dative are compared between native and non-native varieties. If we assume, as the collexeme analyses have shown, that non-native varieties tend to reuse the same lexical items (i.e. types) in the ditransitive dative, type-token ratio should be lower than if they use a diverse set of different lexical items. Indeed, the type-token ratio for recipients in the ditransitive dative is lower in non-native varieties (0.107) compared to native varieties (0.121), and the type-token ratio for themes is lower in the ditransitive dative in non-native varieties (0.272) compared to native varieties (0.310). All in all, results suggest that ditransitive datives are much more concrete and lexically predefined constructions in non-native varieties than in native varieties.

5.6 The register-specificity of the English dative alternation

The results of the mixed-effects model in Section 5.3.4 revealed a third predictor to be regionally malleable: The effect of CORPUS was shown to be weaker in the IndE component of the GloWbE corpus than in data sampled from ICE – this in contrast to all other varieties where no such statistically significant contrast was found. Since the GloWbE data represents one unique register, namely online writing (blogs and general websites), two more specific questions follow, namely (1) does register influence variation in the dative alternation, and if yes, to what extent? And (2) are probabilistic grammars stable across different registers, that is, what is the extent of intra-systemic register-specific variation? The malleability of probabilistic constraints

highlighted in previous analyses will thus be put to the test by a closer investigation of the constraints' cross-lectal and not only cross-regional malleability, that is, the constraints' variability across stylistic lects.

Traditional variationist sociolinguists have long supported the view that grammatical constraints on variation remain stable across different registers (Guy 2005: 562; see also Rickford 2014). As Guy (2005) puts it:

> For the most part, stylistic variation is quantitatively simple, involving raising or lowering the selection frequency of socially sensitive variables without altering other grammatical constraints on variant selection; [...] the grammar is unchanged in stylistic variation. (Guy 2005: 562)

This view has recently been challenged by Grafmiller (2014) who shows that registers vary substantially with regard to the probabilistic constraints that influence syntactic choice (at least in the genitive alternation) (see also Szmrecsanyi 2017a). If we assume that syntactic variation is indeed sensitive to register-specific constraints, as suggested by the cross-corpus differences observed in IndE, a closer look at the register-specificity of the dative alternation is called for.

Such a closer look is especially fruitful if we consider that the present study pools over 14 different registers from informal spoken (dialogue) to written formal (academic writing) including blog and general online data (GloWbE). The aggregate perspective adopted here might thus easily conceal more fine-grained variability within one regional lect. Earlier work on the dative alternation has often taken such a close look at intra-systemic variation within one specific register mainly because the data was sampled from one register only. For instance, Arnold et al. (2000) make use of parliamentary texts to analyse the effect of structural complexity and discourse status. Bernaisch et al. (2014) focus on newspaper language and find that variety as a predictor plays a negligible role concluding that processing-related factors act by and large homogeneously on different types of Englishes. Schilk et al. (2013) also focus on newspaper texts from British English, Indian English and Pakistani English and pinpoint qualitative differences between the three varieties with regard to the predictors that account for variation in the dative alternation. The results in Schilk et al. (2013) highlight that the choice of dative variant is largely shaped by syntactic complexity in all three varieties apart from IndE where recipient pronominality has the largest impact – a finding that is consistent with the present study. Tagliamonte (2014) and Szmrecsanyi et al. (2017) investigate patterns of variation in the dative

alternation in basilectal speech. Tagliamonte reports patterns of harmonic alignment in both Canadian and British English that are taken to the extreme in vernacular speech (Tagliamonte 2014: 314). The latter study by Szmrecsanyi et al. (2017) finds a total of eight statistically significant probabilistic contrasts between four varieties (CanE, BrE, NZE, AmE) that are compared in a pairwise fashion. These contrasts relate to semantics of the verb, recipient pronominality and length. Last but not least, Bresnan & Hay (2008) and Bresnan & Ford (2010) investigate patterns of variation in the dative alternation in telephone conversations and experimental settings and report cross-varietal differences between between American and New Zealand English (Bresnan & Hay 2008) with regard to the effect of recipient animacy, and between American and Australian English (Bresnan & Ford 2010) with regard to length effects. While the aggregate perspective adopted in the present study can partly substantiate the findings of more data-restricted research, especially with regard to the predominant role played by end-weight effects and recipient pronominality in dative grammar, this perspective might also blur possible register-specific variation in the constraints that shape variation. Recall for instance, that recipient animacy was not observed to be cross-regionally malleable in the present analyses which is in contrast to the findings in Bresnan & Hay (2008). Leaving thus the bird's eye perspective, the register-specificity of the dative alternation is investigated subsequently in more detail in two steps to address the two questions outlined above more systematically.

In the first step, the cross-regional variability of register effects is explored by fitting a mixed-effect model on the full dataset with an interaction between variety and register. In addition, a random forest fitted by variety will provide insights into regional variation in the relative ranking of register effects by variety. Second, the internal variability of register-specific probabilistic grammars is explored by fitting a separate mixed-effects model per register and zooming in on the cross-register variability of language-internal constraints.

Five different registers are distinguished in this analysis given by the corpus structure (GENRECOARSE), namely 'dialogue', 'monologue', 'non-printed', 'printed' (registers provided in ICE) and 'online' (the register of GloWbE texts).

5.6.1 Regional variation in register effects

To explore regional variation in the effect of register on the choice of dative variant, a mixed-effects model was fitted to the full dataset. That model includes the five most

important predictors according to the random forest in Section 5.2 as main effects (namely, weight ratio, recipient and theme pronominality, theme head frequency and theme complexity) and the verb, theme, recipient and speaker as random intercepts as well as an interaction between variety and register. Both variety and register are coded with sum coding, again, to compare each level of variety and register against the global mean. Infrequent verbs, themes and recipients occurring fewer than five times were subsumed under 'OTHER'. The model formula is provided in (67). Numeric predictors are again scaled and centred around the mean. Predictions are for the prepositional dative.

(67) Variant ~ (1|VERB) + (1|THEMEHEAD) + (1|RECHEAD) + (1|SPEAKERID) + THEMEHEADFREQ + THEMECOMPLEXITY + RECPRON + THEMEPRON + WEIGHTRATIO + REGISTER + VARIETY + VARIETY:REGISTER

Summary statistics of the model indicate that the model is able to discriminate well between the dative variants (C-statistic = 0.979) and fits the data (accuracy = 93.3%). Model evaluation indicates no sign of overdispersion and moderate collinearity ($\kappa = 13.2$).

Results of the model show that register differs in its effect size statistically significantly in four out of the nine varieties, namely in Hong Kong English, Indian English, Irish English and Jamaican English. Three registers differ in their effect size across those varieties. In sum, four (statistically significant) findings need to be highlighted (see Figure 5.25). First, the prepositional dative is more likely in non-printed texts (light blue dots in Figure 5.25) in HKE, and less likely in IrE compared to the effect of the other registers across varieties. Second, the prepositional dative is more likely in dialogues (light green dots in Figure 5.25) in IndE compared to the effect of the other registers across varieties. Third, the prepositional dative is less likely in online texts (red dots in Figure 5.25) in IndE and more likely in IrE compared to the effect of the other registers across varieties. Finally, the prepositional dative is more likely in printed texts (dark blue dots in Figure 5.25) in JamE and less likely in HKE compared to the effect of the other registers across varieties.

Figure 5.25 visualises the effect of each level of register on the likelihood of a prepositional dative by variety (plotted with the effects package, Fox 2003). The figure highlights that the prepositional dative is the more likely option in most registers in HKE, IndE, JamE, PhiE and SinE (indicated by the points above the dashed 50% line). In the native varieties, on the other hand, namely, CanE, BrE, IrE and NZE, the

ditransitive dative is generally more likely. Only in one register in NZE and BrE is the prepositional dative more likely than the ditransitive dative, namely in monologues in BrE and non-printed texts in NZE.

The effect of GloWbE on dative choice in IndE observed in the first mixed-effects model is thus borne out: Compared to the register effects in all other varieties and the other register effects in IndE, online texts increase the likelihood of a prepositional dative the least in IndE (red dots in Figure 5.25).

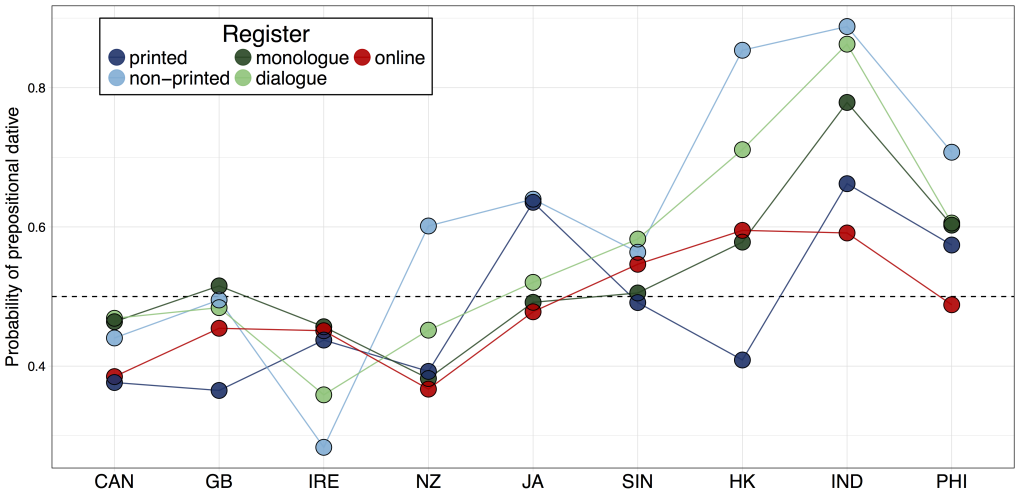


Figure 5.25 The effect of five register levels – printed, non-printed, monologue, dialogue, online – across nine varieties of English based on a mixed-effect model including the five most important predictors — Predictions are for the prepositional dative. Cross-register differences are statistically significant in HKE, IndE, IrE and JamE. Native varieties appear on the left side, non-native varieties appear on the right side of the graph.

In a second step, regional variation in the relative ranking of register was investigated by fitting conditional random forests by variety. To that end, I again made use of the `cforest()` function in the party package and calculated predictor importance with the `varimpAUC()` function. Nine separate forests were fitted with the full range of predictors (see 68). Numeric predictors were again scaled and centred around the mean. Hyperparameters were set to *ntrees* = 3000 and *mtry* = 5.

(68) Variant ~ GenreCoarse + VerbSemantics + WeightRatio + RecGivenness + ThemeGivenness + RecDefiniteness + ThemeDefiniteness + RecHeadFreq + ThemeHeadFreq + RecThematicity + ThemeThematicity + PrimeType + RecPron + ThemePron + RecAnimacy + ThemeAnimacy + TypeTokenRatio

The results of the relative constraint ranking in each variety are presented in Figure 5.26. Constraints are plotted by order of average decreasing importance. Regional differences can be observed in IndE. The relative ranking in IndE is different from the other varieties, not only in ranking recipient pronominality as more important than length but also in assigning register the fifth most important rank. In all other varieties, register has a very marginal impact on dative choice.

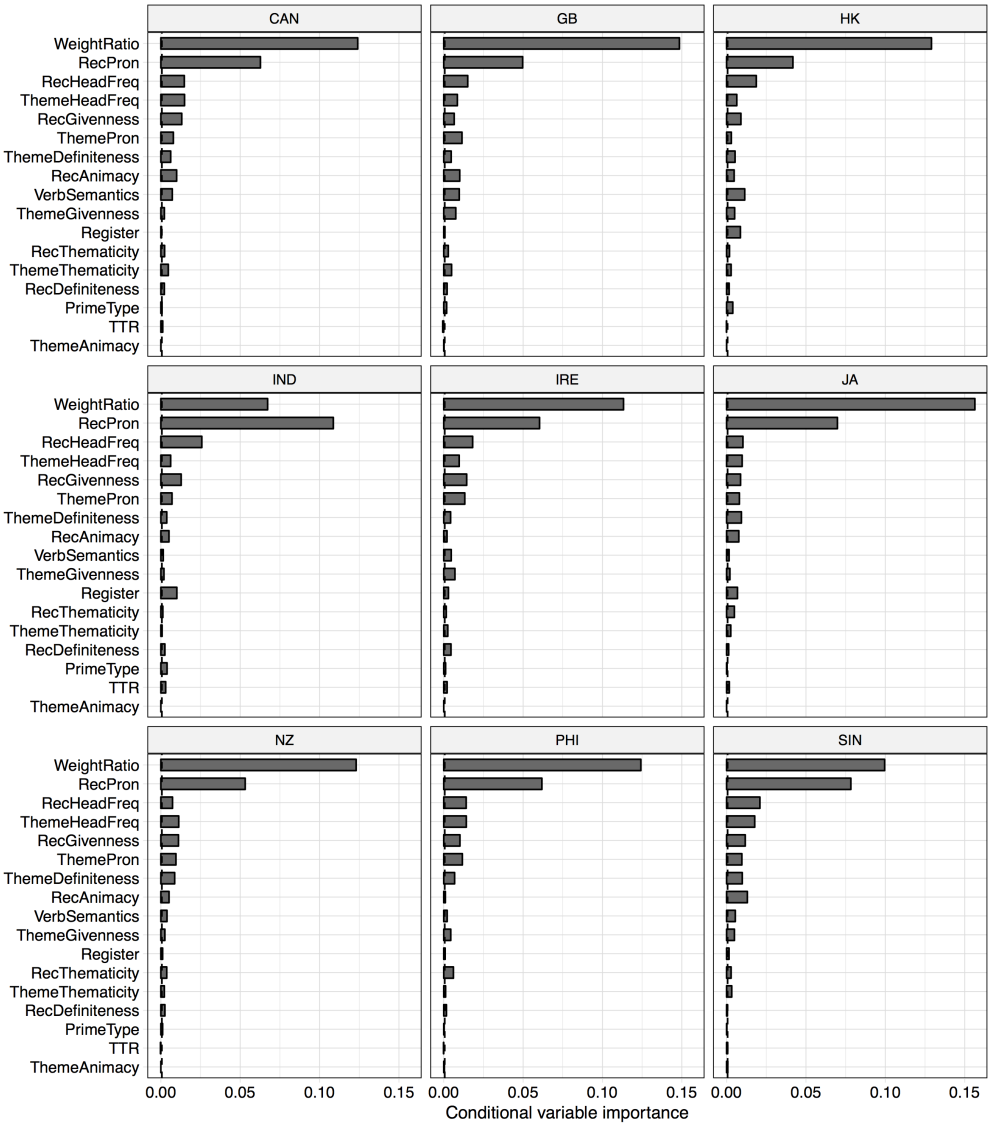


Figure 5.26 Relative ranking of predictors in dative choice with register included — In IndE, register is a fairly important factor (it ranks fifth) in dative choice. In all other varieties, register is only marginally important.

5.6.2 Cross-register variability of probabilistic constraints

The second step of the analysis zooms in on the cross-register variability of language-internal factors in order to assess the extent of intra-systemic register-specific variation. To that end, mixed-effects models were fitted per variety with an interaction between REGISTER and the five most important language-internal factors used in the previous analyses, namely weight ratio, recipient pronominality, theme complexity, theme pronominality and theme head frequency. Again, the random structure of the model includes a random intercept each for recipient, theme and verb. Lexical items occurring eight times or fewer were subsumed under the level 'OTHER'. Note that the threshold is higher than in previous analyses due to model convergence issues. Speaker was excluded as a random intercept due to ensuing problems in model convergence.

Summary statistics of the model indicate overdispersion in the case of the model fitted on JamE but not in any others. Out of the five possible interactions, the effects of recipient pronominality, and to some extent also theme pronominality, were persistently significantly variable across registers (in addition to theme complexity in HKE and IndE and weight ratio in JamE). Recipient pronominality differs in its effect size across registers in four varieties – BrE, IndE, NZE and PhiE – and in three registers, namely online, printed and dialogue.

Changes in the likelihood of prepositional datives with non-pronominal ('non-pron') and pronominal ('pron') recipients are visualised by variety and register in Figure 5.27 with confidence intervals. Statistically significant contrasts are highlighted in grey. The dashed line represents the 50% threshold; points above the line indicate a preference for the prepositional dative, points below the line a preference for the ditransitive dative. Results can be summarised in four main (statistically significant) findings: First, in BrE, the effect of recipient pronominality is weaker in printed texts in comparison to the other registers. Second, the effect of recipient pronominality in IndE is weaker in online texts in comparison to the other varieties (corroborating findings from the earlier mixed-effect model in Section 5.3). At the same time, the effect is stronger in dialogues in IndE. Third, the effect of recipient pronominality in NZE is weaker in online texts and stronger in printed texts. And fourth, in PhiE, the effect of recipient pronominality is weaker in online texts compared to other registers. Recall that the observed differences always have to be considered in contrast to the effect of the other registers within the same variety.

Changes in the likelihood of a prepositional dative variant with non-pronominal

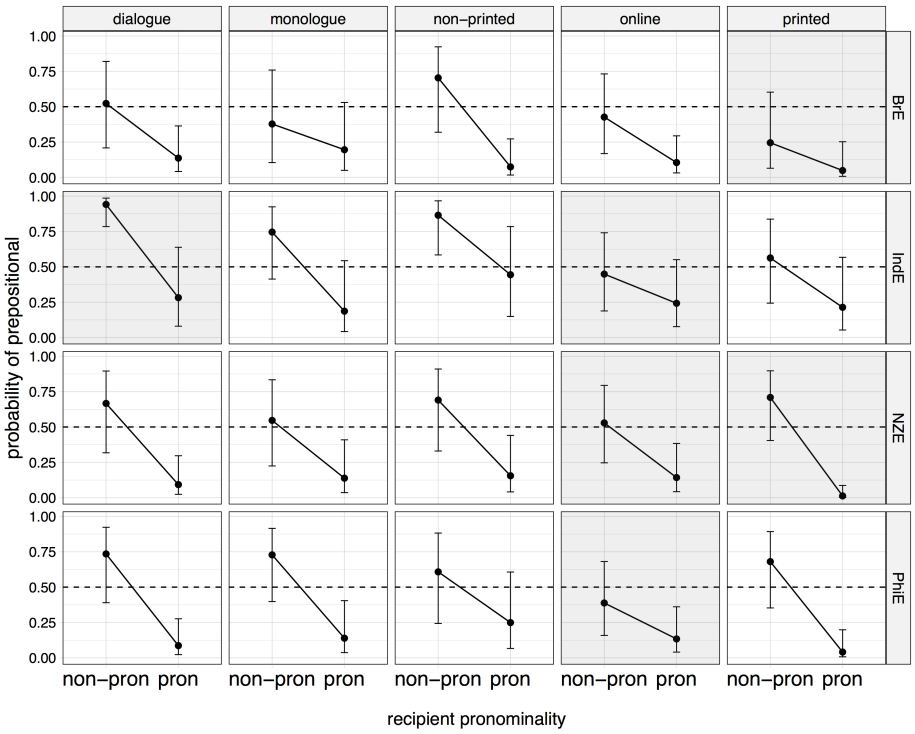


Figure 5.27 The effect of recipient pronominality across five registers and four varieties with confidence intervals — Statistically significant contrasts are highlighted in grey. The dashed line represents the 50% threshold in the likelihood of a prepositional dative.

('non-pron') and pronominal ('pron') themes are visualised by the two varieties and the two registers in Figure 5.28 with confidence intervals. Statistically significant contrasts are highlighted in grey. The dashed line represents the 50% threshold; points above the line indicate a preference for the prepositional dative, points below the line a preference for the ditransitive dative. Theme pronominality differs in its effect size across registers in two varieties – BrE and JamE – and in two registers, namely printed texts and dialogues. In BrE dialogues, the effect of theme pronominality is stronger compared to all other registers. At the same time, in both BrE and JamE, theme pronominality has the opposite effect in printed texts compared to what is expected on a global scale: The prepositional dative (in which the theme is the first constituent) is less likely if the theme is pronominal and more likely if it is non-pronominal. This reverse effect contradicts all findings of earlier studies that observed an increase in the likelihood of a prepositional dative with pronominal themes and thus a congruence of the effect of theme pronominality with other predictors: Pronominal themes commonly increase the likelihood of a prepositional dative and not the other way

round. This disrupted harmonic alignment effect of theme pronominality in dative choice suggests that accessibility or ease of processing of the theme might not be as important in printed texts in BrE and JamE as in other registers. Taking a closer look at the data reveals that observations with pronominal themes in ditransitive datives and non-pronominal themes in prepositional datives are very sparsely distributed in the printed texts. Ditransitive datives with pronominal themes from BrE and JamE often include impersonal pronouns, such as *anything, nothing* or *something*.

(69) *The taxpayer* ***owes Sheffield nothing*** ... <ICE-GB:W2E-005>

(70) ... *Charlotte* ***owed my grandmother something*** ... <ICE-JA:W2F-013>

Prepositional datives with non-pronominal themes are even rarer. The prepositional dative is probably more likely in those contexts due to a verb bias towards the prepositional dative (e.g. with *extend, pay, set*).

(71) *If I'd* ***set my mind to it****.* <ICE-GB:W2F-010>

(72) ... *to* ***pay his respects to her*** *the minute he arrived.* <ICE-JA:W2F-005>

In light of these findings, any conclusion about the cross-register malleability of theme pronominality needs to be tentative, especially since the large confidence intervals in Figure 5.28 are indicative of sparse data. In contrast, the cross-register variability of the effect of recipient pronominality seems robust pointing to the cross-lectal variability of this constraint. Out of all constraints then, recipient pronominality turns out to be the one constraint that is variable inter- and intra-systemically.

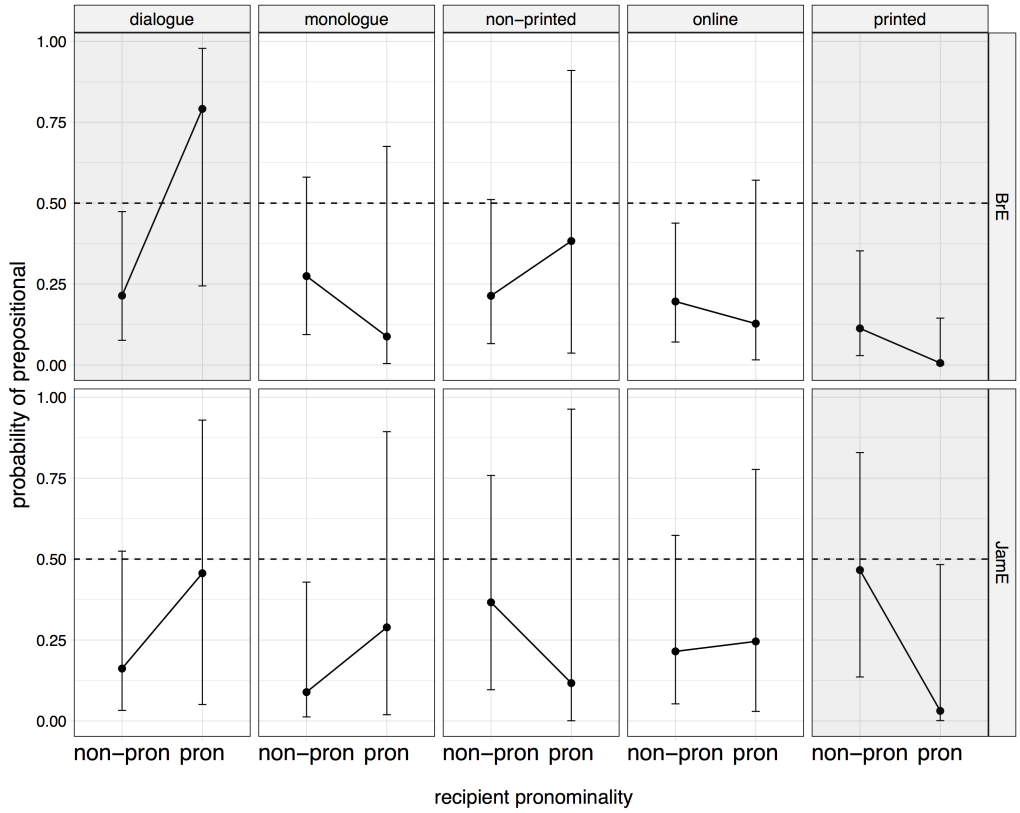


Figure 5.28 The effect of theme pronominality across five registers and two varieties with confidence intervals — Statistically significant contrasts are highlighted in grey. The dashed line represents the 50% threshold in the likelihood of a prepositional dative.

5.6.3 Interim summary

The results of the two analyses presented above indicate that register influences syntactic variation in the dative alternation differently across varieties. The effect of register is statistically significant in HKE, IndE, IrE and JamE. The effect of all registers is thereby statistically significant with the exception of 'monologue'. What is more, the first analysis reveals that most registers in non-native varieties make the prepositional dative more likely (in HKE, IndE, JamE, PhiE and SinE). In contrast, register effects in native varieties decrease the likelihood of a prepositional dative below the 50% threshold. The second analysis has furthermore shown that probabilistic constraints are not only regionally but also stylistically malleable. First and foremost the effect of recipient pronominality displays intra-systemic variability.

The present study had set out to investigate cross-varietal variation of register in the English dative alternation as a consequence of the statistically significant effect of

CORPUS in IndE. The two analyses illustrate, however, that the probabilistic grammar of other varieties, such as IrE, JamE and HKE, are also register-specific and in that regard statistically significantly different from the global norm.

While the cross-lectal malleability of factors constraining the dative alternation has now received ample attention, the extent to which probabilistic grammars of varieties of English can be said to be similar to or different from each other still needs to be determined. In order to do so, the next section will draw on methods developed in comparative sociolinguistics which are especially suited to provide a comparative perspective on the probabilistic constraints that impact variation.

5.7 Assessing the stability of probabilistic grammars

Probabilistic accounts share their interest in the gradient variability in quantitative patterns not only with the developing field of Cognitive Sociolinguistics (see, for instance Geeraerts et al. 2010) but their research agenda and methodology also have a long-standing tradition in Variationist Sociolinguistics (e.g. Labov 1972b, 1982; see also Chapter 1). It is especially one subfield of Variationist Sociolinguistics, namely Comparative Sociolinguistics, whose comparative perspective is particularly relevant for the present exploration of the limits of cross-varietal variation and hence the stability in probabilistic grammars (Poplack & Tagliamonte 2001; Tagliamonte 2012, 2002; and many others). Comparative Sociolinguistics makes use of the methods developed in Variationist Sociolinguistics, that is, the disentangling of various social and cognitive constraints and their influence on linguistic features using multivariate analysis, and compares and contrasts these patterns of variability of linguistic features across different dialects or varieties in order to approximate a common source of shared dialect features. The results from a quantitative variationist approach are thereby essential to identify differences and/or similarities between varieties and dialects (Tagliamonte 2012: 164; see also Poplack & Tagliamonte 2001). Similarities and differences are identified along three dimensions, also called the "three lines of evidence" (Poplack & Tagliamonte 2001: 92; Tagliamonte 2002: 731). The first line compares varieties based on the statistical significance of predictors by determining which factors are statistically significant below or at the $p = .05$ threshold for each variety. The second line compares varieties based on effect size or relative strength of factors. The third line takes the constraint hierarchy in each variety into account, that is, the ranking of factors that constrain a linguistic variable (Tagliamonte 2012: 122).

By comparing patterns of variability along these three lines of evidence across different dialects or varieties, we can identify so-called *conflict sites* which are defined as the "form[s] or class[es] of forms that differ[] functionally and/or structurally and/or quantitatively across the varieties in question" (Tagliamonte 2012: 164). At the same time, linguistic similarity between varieties can be determined on the grounds of shared features from each of these three dimensions (Tagliamonte 2012: 166).

Hence, one way to assess the overall stability of probabilistic grammars and thus the limits of cross-varietal variation across different speech communities is to apply the above outlined comparative method to the probabilistic grammars underlying the dative alternation. Since previous analyses have shown that stability largely prevails in probabilistic constraints (Bernaisch et al. 2014), we expect overall a high degree of similarity between the varieties on all three levels of comparison. Further, dissimilarities in probabilistic grammars are expected for those varieties where we could observe statistically significant deviations from the global average.

In order to compare probabilistic grammars along three different dimensions, separate mixed-effect models (for the first and second line) and separate random forests (for the third line) were fitted to each variety. Due to convergence and model fitting issues (and for ease of comparability with similar analyses, see Heller 2018), attention was restricted to the five most important predictors from the random forest analysis (shown in Section 5.2.2). The final predictors include weight ratio, recipient pronominality, theme pronominality, theme complexity and theme head frequency. The random structure was restricted to a random intercept for verb, the theme head, the recipient head and for file number in order to make computation more feasible. Note that the model for Irish English did not converge with speaker as a random intercept, hence the use of file number. When fitting the models, infrequent lexical items in random effects (verb, theme, recipient) occurring five times or fewer were grouped under 'OTHER'. Again, this made computation for all subsets more efficient. Numeric predictors (weight ratio and theme head frequency) were centred around the mean and scaled by two standard deviations. Categorical predictors were also centred around their mean (by turning them into numeric values). The model formula is given in (73).

(73) Variant ~ (1|VERB) + (1|THEMEHEAD) + (1|RECHEAD) + (1|FILEID) + WEIGHTRATIO + RECPRON + THEMEPRON + THEMECOMPLEXITY + THEMEHEADFREQ

Model predictions are again for the prepositional dative. Summary statistics indicate

that all models are well able to discriminate between the two dative variants (the lowest *C*-statistic is 0.959 in British English) and also fit the data (lowest accuracy is 90.7% in Singapore English).

A comparative sociolinguistic approach to modelling the stability of probabilistic grammars entails the comparison between the output of the nine models with regard to the statistical significance of predictors (is a predictor significant or not), the relative strength of predictors (coefficient estimates) and the constraint hierarchy (ranking of predictors). For this reason, different similarity and dissimilarity measures were used depending on the data to be compared.

Since a comparison among nine varieties leaves us with a 9×9 dimensional distance matrix, the number of dimensions had to be reduced for visualisation purposes. To that end, the three comparisons make use of multidimensional scaling (MDS) (Kruskal & Wish 1978) using the `cmdscale()` function in R for classical metric MDS (R Core Team 2016) and `isoMDS()` from the MASS package for non-metric MDS (Venables & Ripley 2002). Recall that, MDS creates another data frame approximating the original input data from which another distance matrix is calculated with a reduced number of dimensions. A number of iterations are performed to calculate this approximate data frame and distance matrix until the difference between the original distance matrix and the recalculated distance matrix is as low as possible. The difference between the two distance matrices, called *stress*, provides an indication of goodness-of-fit. Stress ranges between 0 and 1 with zero indicating perfect fit and 1 indicating random noise and no fit at all. Besides visualising the probabilistic distance between varieties and thus the stability of probabilistic grammars, a mean similarity score can be computed based on the calculated distance/similarity matrices. This mean similarity score provides an indication of the stability of probabilistic grammars with regard to that specific measurement (significance, relative strength, ranking). As will be shown, the three comparative measures show hardly any convergence in their distance measures.

In sum, the following steps will be taken for the three lines of comparison and applied to an investigation of inter-systemic stability in the dative alternation in the subsequent sections:

1. Calculate/Create a distance matrix:
    - **statistical significance:** compare number of shared significant and non-significant predictors
    - **relative strength:** use Euclidean distance metric to calculate distance

between coefficient estimates from models

- **constraint ranking:** calculate Spearman's rank correlation coefficient between the constraint ranks as a distance measure

2. Calculate the average similarity as a measure of overall stability

3. Reduce number of dimensions of distance matrix:

   - **statistical significance:** non-metric MDS
   - **relative strength:** classical metric MDS
   - **constraint ranking:** non-metric MDS

4. Plot the dimensions

5.7.1 Comparing statistical significance

Out of the five factors included in the by-variety models, only relative length and recipient pronominality are statistically significant in all nine varieties. This finding is consonant with the results of previous analyses that highlighted the overarching importance of these two predictors. The remaining three factors are alternately significant or not significant across varieties without any clear patterns emerging (see Table 5.17).

Table 5.17 Significant and non-significant predictors in nine varieties of English based on mixed-effects logistic regression — Plus (+) indicates that the predictor is significant at $p<.05$; minus (-) indicates non-significance of that predictor in that particular variety.

Factor	CanE	BrE	HKE	IndE	IrE	JamE	NZE	PhiE	SinE
weight ratio	+	+	+	+	+	+	+	+	+
recipient pronominality	+	+	+	+	+	+	+	+	+
theme complexity	–	+	+	+	+	–	+	–	+
theme pronominality	+	–	–	–	–	–	–	+	+
theme head frequency	+	–	–	–	–	–	–	–	–

In order to calculate the differences between varieties based on the presence or absence of significant predictors, the number of shared and not-shared significant predictors were calculated in a pairwise comparative fashion. For instance, comparing CanE and BrE, we find that the two varieties share significant predictors in two cases (namely weight ratio and recipient pronominality) and they are dissimilar with regard

to the other three predictors (for instance, theme complexity is significant in BrE but not in CanE). Hence, the comparison between British and Canadian English receives a similarity score of 2 (out of 5) and a dissimilarity score of 3 (out of 5). Comparing Indian and Irish English, we find an overlap of three significant and two non-significant predictors. The comparison between Indian and Irish English receives therefore a similarity score of (3+2=) 5 and a dissimilarity score of 0. This dissimilarity or distance score corresponds to the Manhattan or City-Block distance between two objects (or varieties in this case), which is a special case of Euclidean distance (Aldenderfer & Blashfield 1984: 25) (see Section 4.4 on the difference between Manhattan and Euclidean distance). If two varieties agree on the significance of all five predictors, their comparison is assigned a similarity score of 5 and a dissimilarity score or Manhattan distance of 0. If two varieties do not agree with regard to the significance of all five predictors, their comparison is assigned a similarity score of 0 and a Manhattan distance of 5. Since we are interested in the similarity between varieties (as a measure of overall stability of probabilistic grammars), let us focus on similarity for the moment.

Comparing the number of shared significant and non-significant predictors across all nine varieties results in 36 similarity scores which are displayed in a similarity matrix in Table 5.18. Low numbers indicate low similarity between varieties, high numbers (up to 5) indicate predominant similarity between varieties.

Table 5.18 Similarity matrix based on statistical significance of five predictors in per-variety mixed-effects logistic regression — Similarity scores range between 0 and 5 with 0 indicating no similarity at all and 5 complete similarity.

	CanE	BrE	HKE	IndE	IrE	JamE	NZE	PhiE
BrE	2							
HKE	2	5						
IndE	2	5	5					
IrE	2	5	5	5				
JamE	3	4	4	4	4			
NZE	2	5	5	5	5	4		
PhiE	4	3	3	3	3	4	3	
SinE	3	4	4	4	4	3	4	4

Next, the similarity scores from this similarity matrix were scaled by dividing them by the maximum similarity score (5) to get values that fit in the range between 0 and

1. The similarity score between CanE and BrE of 2 was consequently transformed to $2/5 = 0.4$; the similarity score between IndE and IrE of 5 was transformed to 1. The mean of these scaled similarity scores was then calculated for each variety. Mean similarity is highest for BrE, IndE and JamE, and lowest for IrE and CanE, that is, IrE is least similar to all other varieties (see Table 5.19).

Table 5.19 Mean similarity of nine varieties of English based on the number of shared significance — Values towards 1 indicate absolute similarity (with all other varieties), values towards 0 indicate absolute dissimilarity. Varieties are ordered by decreasing size of mean similarity.

Variety	Mean similarity
BrE	0.825
HKE	0.825
IndE	0.825
IrE	0.825
NZE	0.825
JamE	0.750
SinE	0.750
PhiE	0.675
CanE	0.500

From a bird's eye perspective, we observe that the North American (influenced) varieties, that is, Philippine and Canadian English (see Section 3.10 on the history of PhiE), are the least similar to all other varieties, while British English is the most similar to all other varieties. In total, the average similarity across all varieties sums up to a stability score of 0.756.

To visualise the probabilistic distance between the varieties, the similarity matrix (Table 5.18) was transformed ino a dissimilarity matrix (by subtracting the similarity scores from the number of possible overlaps, i.e. 5) which was then used as input for a multidimensional scaling analysis.

Recall that MDS is a dimension reduction technique, so in essence the 9×9 distance or dissimilarity matrix is reduced to a two-or-three-dimensional matrix where the distances between varieties approximate the distances in the original 9×9 distance matrix. Since the distance matrix is based on non-metric measurements for the current analysis, that is, logical values, `isoMDS` was used to reduce the number of dimensions (stress of the MDS solution = 0.007, which is excellent). The MDS plot (Figure 5.29)

shows that BrE, IndE, NZE, HKE and IrE are identical with regard to the number of shared significant and non-significant predictors. On the opposite side of the plot on the first (horizontal) dimension, PhiE and CanE cluster close to each other. On the second (vertical) dimension, SinE and JamE are on opposite ends. Distances between varieties correspond to differences in shared significant and non-significant predictors.

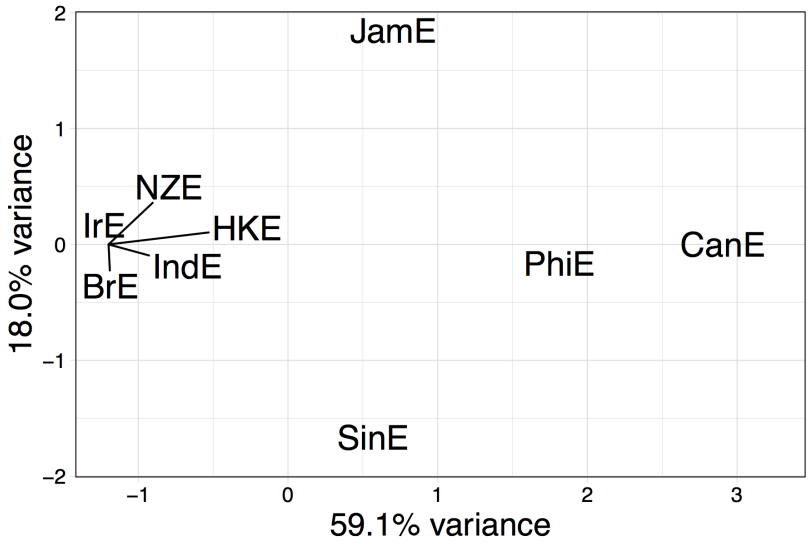


Figure 5.29 Multidimensional scaling map of nine varieties of English — Distances between varieties correspond to differences in the number of shared non-significant and significant predictors.

In sum, the first comparison shows that CanE and PhiE are the least similar to the other varieties, and that NZE, BrE, HKE, IndE and IrE are identical with regard to significant and non-significant predictors.

5.7.2 Comparing relative strength of predictors

For the second comparative line of evidence, the difference and similarity between probabilistic grammars was assessed based on the coefficient estimates obtained from the same per-variety mixed-effects models used before (with the five most important predictors). Coefficient estimates give an indication of the strength of the predictor in the choice of dative variant. To that end, the models' coefficient estimates were extracted (without the intercept) into a separate data frame (see Table 5.20).

Next, Euclidean distance was calculated between the varieties in a pairwise fashion on the basis of the coefficient estimates using the `dist()` function from the stats package (R Core Team 2016). Calculating the Euclidean distance between the

Table 5.20 Coefficient estimates from per-variety mixed-effects models — Predictions are for the prepositional dative.

Factor	**CanE**	**BrE**	**HKE**	**IndE**	**IrE**	**JamE**	**NZE**	**PhiE**	**SinE**
weight ratio	4.06	3.23	3.25	2.83	2.90	5.33	2.80	4.31	2.35
recipient pronom.									
pron ⇒ non-pron	3.35	1.95	2.49	2.79	3.00	2.53	2.75	3.12	2.78
theme complexity									
complex ⇒ simple	0.20	0.81	0.77	1.16	0.96	0.28	0.82	0.69	1.26
theme pronom.									
non-pron ⇒ pron	2.00	-0.38	0.72	1.09	1.22	-0.42	1.36	2.33	1.60
theme head freq.	1.13	0.63	0.34	0.14	-0.01	-0.77	0.48	0.36	-0.01

nine varieties of English resulted in a 9×9 dimensional distance matrix (shown in Table 5.21).

Table 5.21 Euclidean distances between nine varieties of English based on coefficient estimates from per-variety mixed-effects models

	CanE	BrE	HKE	IndE	IrE	JamE	NZE	PhiE
BrE	2.991							
HKE	1.999	1.256						
IndE	2.140	1.837	0.773					
IrE	1.997	2.046	0.893	0.353				
JamE	3.428	2.644	2.662	3.189	3.141			
NZE	1.790	1.964	0.839	0.555	0.594	3.387		
PhiE	1.028	3.151	2.026	2.027	1.862	3.220	1.843	
SinE	2.417	2.447	1.429	0.729	0.767	3.818	0.827	2.223

In order to be able to compare the results of the second line of evidence to the stability score of the first line of evidence, similarity scores had to be calculated from these distances. This presented a challenging task: Euclidean distance measures have a lower boundary in that 0 equals no distance, but no upper boundary in absolute terms (Aldenderfer & Blashfield 1984: 27). The distance values always depend on the input data received. For instance, a distance of 2.991 (between, for instance, British and Canadian English) might be very small if we assume that the maximum distance

between two varieties is 100. However, if the maximum distance is 4, then British and Canadian English would in fact be very distinct from each other. To overcome this problem and to find the maximum distance possible, a null-model was added to the equation. This null-model includes coefficient estimates that all have the value of 0, i.e. none of the predictors have any effect. The coefficient estimates of the null-model (all zeros) were added to the data frame in Table 5.20 and the Euclidean distance was calculated again among all ten models. The new distance matrix includes a tenth variety now, called NULL, and its distance in Euclidean terms from the other nine varieties. The mean of all distances was then calculated for each variety. Mean distance is (obviously) highest for the null-model and lowest for NZE and IndE, that is, IndE is most similar to all other varieties (see Table 5.22).

Table 5.22 Mean distances of nine varieties of English and a null-model based on coefficient estimates from per-variety mixed effects models — Values towards 0 indicate absolute similarity (with all other varieties), values towards the highest value (the null-model) indicate absolute dissimilarity.

Variety	Mean distance
NULL	4.768
JamE	3.495
CanE	2.616
PhiE	2.582
BrE	2.473
SinE	2.092
HKE	1.791
IrE	1.789
NZE	1.784
IndE	1.765

Having an upper boundary of maximum distance (i.e. the null-model) allows for the normalisation of all mean distance scores and their transformation into similarity scores. To normalise the values, each original mean distance score (calculated without the null-model) was divided by the mean distance of the null-model (for instance, in JamE: 3.186/4.768) and then subtracted from 1 to arrive at a similarity score (shown in Table 5.23). The mean similarity across all varieties (i.e. the stability score for the second line of evidence) is 0.591 and thus lower than the stability score of the first line of evidence (0.756).

Table 5.23 Mean similarities across nine varieties of English based on coefficient estimates from per-variety mixed effects models — Values towards 0 indicate absolute dissimilarity (with all other varieties), values towards 1 indicate absolute similarity.

Variety	Mean similarity
IndE	0.696
IrE	0.694
NZE	0.691
HKE	0.689
SinE	0.616
PhiE	0.544
CanE	0.534
BrE	0.519
JamE	0.332

To visualise the similarity among varieties, the same steps were taken as in the previous section and the distance matrix in (5.21) was reduced to a lower dimensional space for the purpose of interpretation and visualisation, using the `cmdscale()` function from the stats package (R Core Team 2016). Two dimensions account for 89.0% of the variance in the 9×9 dimensional matrix (stress = 0.113, which is fair).

The first dimension, which accounts for 53.3% of the variance in the original distance matrix, splits JamE from the rest of the varieties, followed by BrE. The second dimension (accounting for 35.8% of the variance) further splits CanE and PhiE from IrE, IndE, SinE, NZE, HKE and BrE and also increases the divide between JamE and the other varieties. All in all, CanE and PhiE form a cluster of American English (influenced) varieties, IrE, IndE, SinE, NZE, HKE and BrE are fairly close to each other and JamE is plotted all by itself (see Figure 5.30).

In sum, the second line of evidence indicates that IndE is the most similar and JamE the most dissimilar from all varieties. This dissimilarity is visualised in the MDS plot where JamE is plotted away from the rest. The MDS further plots CanE and PhiE together suggesting a cluster of American-based varieties. Mean similarity score is 0.591 which is lower than the similarity score from the first comparison.

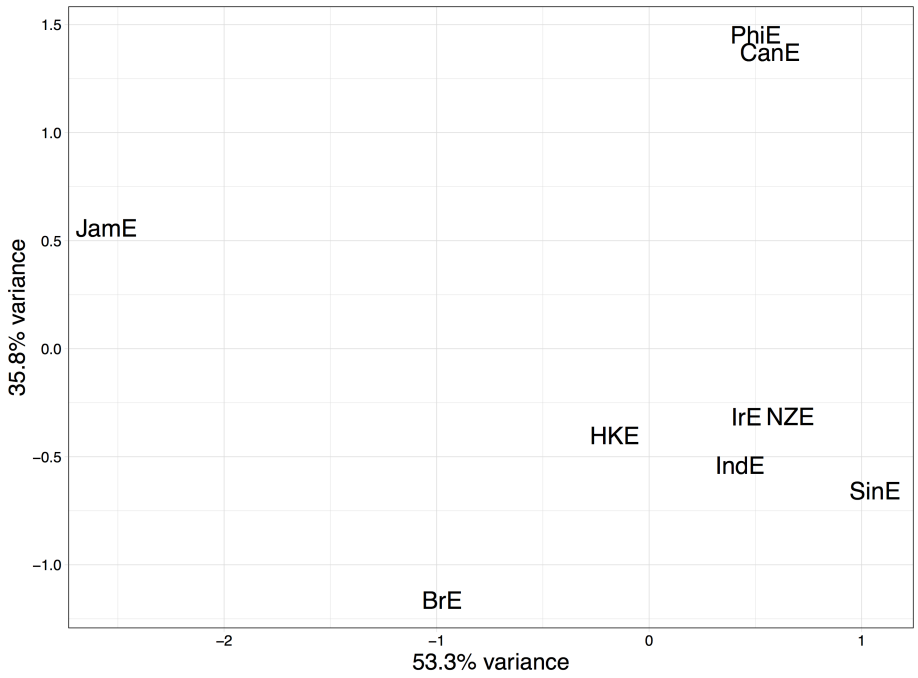


Figure 5.30 Multidimensional scaling map of nine varieties of English — Distances between varieties correspond to differences in the coefficient estimates of the five most important predictors. Each axis label indicates the amount of variance accounted for by the dimension.

5.7.3 Comparing ranking of variable importance

Finally, the last line of evidence compares the constraint ranking of predictors across varieties in order to assess cross-regional stability. For that purpose, separate random forests were fitted by variety to obtain the constraints' relative importance. Conditional random forests were computed using the `cforest()` function in the party package (Hothorn et al. 2006a; Strobl et al. 2007, 2008) and included the same five predictors that were used in the per-variety mixed-effects models. The model formula is repeated in (74) (see also 73). Variable importance was calculated with the `varimpAUC()` function in the party package (Janitza et al. 2013). Parameters were set to *mtry* = 3 and *ntree* = 2000 (see also Chapter 4.4 for more details on the technique).

(74) Variant ~ WEIGHTRATIO + RECPRON + THEMEPRON + THEMECOMPLEXITY + THEMEHEADFREQ

Table 5.24 shows the rank of each factor with 1 indicating the most important and 5 the least important factor. Comparison across the nine varieties reveals that weight ratio

and recipient pronominality are the two most important predictors across all varieties. Only in IndE is the ranking of these two factors reversed (shown in Table 5.24). Note that fluctuations for the remaining three predictors can range across all three ranks. For instance, theme head frequency, which ranks third on a global scale (averaged across all random forests), is ranked third, fourth or fifth in individual varieties.

Table 5.24 Factor rankings in per-variety random forests — Factors are ordered by the average global ranking.

Factor	**CanE**	**BrE**	**HKE**	**IndE**	**IrE**	**JamE**	**NZE**	**PhiE**	**SinE**
weight ratio	1	1	1	2	1	1	1	1	1
recipient pronominality	2	2	2	1	2	2	2	2	2
theme head frequency	3	4	3	5	3	3	4	3	3
theme complexity	5	3	4	3	4	4	3	5	4
theme pronominality	4	5	5	4	5	5	5	4	5

The ranking of the five predictors in each random forest was then compared quantitatively using Spearman's rank correlation coefficient (also known as ρ or *rho*). Spearman's rank correlation coefficient calculates the correlation between two vectors (here: variety-based vectors) based on the ranks of the observations in the two vectors (Baayen 2008: 91). Its values range from -1 (negative correlation) to 0 (no correlation) to 1 (positive correlation); the larger the correlation coefficient in absolute numbers, the more similar the two vectors. The correlation was computed in a pairwise fashion using the *cor.test()* function from the stats package (R Core Team 2016), which resulted in a similarity matrix comparable to the previous ones (shown in Table 5.25). In contrast to the previous two comparisons, Spearman's rank correlation coefficient was used as direct input for probabilistic distances between varieties instead of an traditional distance metric such as Manhattan or Euclidean distance measures.

Next, the mean correlation per variety was calculated to obtain mean similarity per variety (see Table 5.26). No clear-cut distinction between varieties on socio-historical grounds can be observed. Rather, HKE shows the highest degree of mean similarity together with JamE and IrE. IndE exhibits the lowest mean similarity in a global comparison, together with CanE and PhiE. Overall, global mean correlation amounts to 0.839, suggesting a comparably high level of stability in predictor ranking across varieties (see Heller 2018).

Table 5.25 Spearman's rank correlation coefficient per variety-pair — 0 is the lowest, 1 the highest value of similarity.

	CanE	BrE	HKE	IndE	IrE	JamE	NZE	PhiE
BrE	0.7							
HKE	0.9	0.9						
IndE	0.5	0.8	0.6					
IrE	0.9	0.9	1.0	0.6				
JamE	0.9	0.9	1.0	0.6	1.0			
NZE	0.7	1.0	0.9	0.8	0.9	0.9		
PhiE	1.0	0.7	0.9	0.5	0.9	0.9	0.7	
SinE	0.9	0.9	1.0	0.6	1.0	1.0	0.9	0.9

Table 5.26 Mean similarities in nine varieties of English based on predictor rankings from per-variety random forests — Values towards 0 indicate absolute dissimilarity (with all other varieties), values towards 1 indicate absolute similarity.

Variety	**Mean correlation**
HKE	0.900
IrE	0.900
JamE	0.900
SinE	0.900
BrE	0.850
NZE	0.850
CanE	0.813
PhiE	0.813
IndE	0.625

In order to visualise the probabilistic distances between varieties, multidimensional scaling was once more employed using the `isoMDS()` function to arrive at an interpretable lower dimensional matrix. Similar to the first line of evidence, the input is again non-metric. Since multidimensional scaling requires a distance and not a similarity matrix as input, the correlation coefficients were transformed to distance measures by subtracting the absolute value of the coefficients from 1.

The MDS solution (stress = 0.007) groups BrE and NZE together. SinE, IrE, JamE and HKE all form one straight clustered line, CanE and PhiE overlap and IndE is plotted away from these three variety clusters. While no native versus non-native cluster is observed, we nevertheless again find a cluster of North American (influenced) varieties

(CanE, PhiE), a fuzzy cluster of British native varieties (BrE, NZE) and a cluster of non-native varieties (without IndE but including IrE) (shown in Figure 5.31).

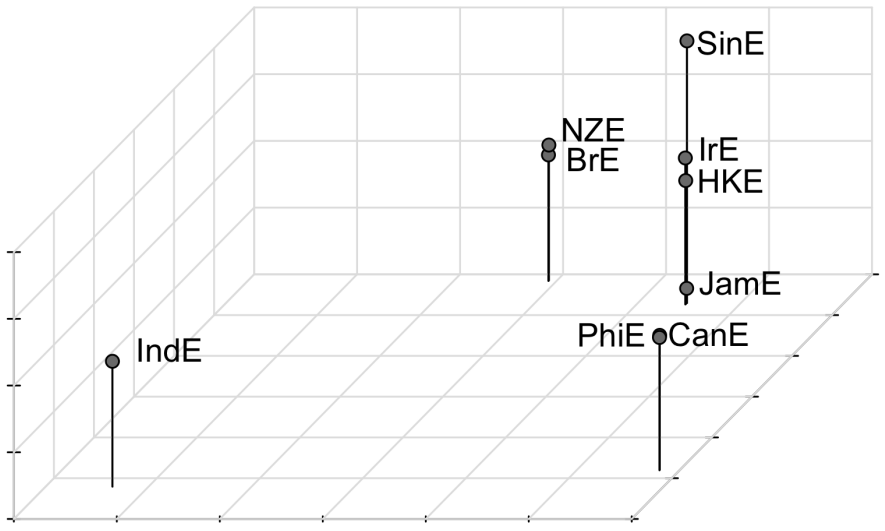


Figure 5.31 Multidimensional scaling map of nine varieties of English — Distances between varieties correspond to differences in the ranking of predictors between by-variety conditional random forests.

The mean similarities from all three comparisons are summarised in Table 5.27. The overall stability score across all three lines of evidence averages to 0.729.

Table 5.27 Stability scores across three lines of evidence

Line of evidence	**Mean similarity score**
1st line: significance	0.756
2nd line: effect size	0.591
3rd line: constraint hierarchy	0.839
average stability score	0.729

5.7.4 Comparing stability scores within and across

Taken together, the three stability scores point to a lower stability of the probabilistic grammar underlying the English dative alternation than the genitive alternation (shown in Table 5.28).

What is more, the variety clusters that emerge in the MDS solutions cannot always

Table 5.28 Stability scores across three lines of evidence in comparison — Overall, the genitive alternation is probabilistically more stable than the dative alternation.

Line of evidence	Mean datives	Mean genitives
1st line: significance	0.756	0.726
2nd line: effect size	0.591	0.804
3rd line: constraint hierarchy	0.839	0.958
average stability score	0.729	0.829

be explained on socio-historical grounds. What we do find in all three comparisons is a cluster of North American (influenced) varieties, namely CanE and PhiE. In both varieties, American (US) English constitutes the historical input variety. In two out of three comparisons, Jamaican English is plotted away from the other varieties (1st and 2nd line of evidence). No other pervasive patterns could be discerned. It thus seems as if the three lines of evidence highlight different patterns of similarity and dissimilarity and measure different aspects of speakers' probabilistic grammar which are not necessarily compatible with each other. This incompatibility is empirically supported by a mantel test (using `mantel()` function in the vegan package, see Oksanen et al. 2017) which was used to calculate the correlation between the three distance matrices. Pearson's moment correlation coefficient between the matrices of the first and second line of evidence is $r = 0.4582$ ($p = .067$), between the first and the third line it is $r = 0.2242$ ($p = .176$) and between the second and the third line it is $r =$ -0.09635 ($p = .568$). In sum, none of the matrices are significantly correlated and the correlation is small (0 indicates no correlation, ±1 indicates complete correlation). Even the comparison of all three matrices (with `mantel.partial()`) indicates no significant correlation ($r = 0.4947$, $p = .05$).

This disparity could call into question the suitability of the comparative sociolinguistic methods applied here. However, a simulation study on the genitive alternation (see also Heller 2018) and a bootstrapping with random data subsampling confirm the appropriateness and the reliability of the three comparisons.

In his simulation study, Heller (2018) increases the amount of variability in the data by allowing the coefficient estimates more range in standard deviation in incremental steps and compares the nine varieties along the three dimensions when variability increases. All three lines pick up on increasing variability, that is, the similarity between the varieties decreases when the coefficient estimates are allowed

to range more widely (see also Appendix B for a more detailed description). The results of the simulation study also show that combining the three lines of evidence gives the best estimate of overall stability. The concept of the three lines of evidence can thus be validated. But how reliable are the methods?

In order to test concept reliability, I ran a bootstrap on my own dataset, fitting 1,000 models on a random selection of 50% of the data for each variety, again using the same set of five predictors for both the regression models (1st line and 2nd line) and the random forests (3rd line). The confidence intervals of the 1,000 models can then be used as an indication of the reliability of each measure. The assumption is that the more reliable the measure, the more it should be independent from the (randomly) sampled data and the narrower the confidence intervals should be around the mean. Results of that bootstrapping show that relative effect strength (i.e. the coefficient estimates) is the most reliable measure. However, coefficient estimates also generate the lowest stability score (see Figure 7 in Appendix B). Note that due to computational limitations, only 50 runs of the random forests were fitted.

Finally, a caveat remains to be added: While the three MDS solutions presented above provide insightful aspects about the probabilistic distance and the stability of probabilistic grammars worldwide, it has to be kept in mind that they only take one single (syntactic) variable into account. What is more, the way this variable was used to measure distances deviates from the general approach in linguistics where distances between varieties or dialects are measured based on the frequency of occurrence or the absence/presence of linguistic features, rather than on underlying probabilities such as the ranking of predictors, the level of significance and effect sizes.

Since this study adopts the usage-based perspective of probabilistic grammars, which posits that statistical regularities (the probabilities measured in the three lines of evidence) are derived from repeated exposure to linguistic items (the morphosyntactic features explored in traditional dialectometric research), it is only appropriate to compare the probabilistic distance between varieties with their morphosyntactic distance derived from traditional measures. If we assume that variability in probabilistic grammars is a reflection of the morphosyntactic variability observed in these varieties, we can hypothesise that distances calculated on the basis of probabilistic grammars and on morphosyntax should be fairly similar. To further explore this empirically, the results of the three lines of evidence were correlated with distance measures obtained on the basis of the feature catalogue accompanying the *Electronic World Atlas of Varieties of English* (Kortmann & Lunkenheimer 2013). This catalogue includes

information on the presence and absence of 76 morphosyntactic features from 46 vernacular varieties of English, using a four-scaled measure of presence: 'A' indicates that the feature is pervasive; 'B' indicates that the feature is neither pervasive nor extremely rare; 'C' means that the feature exists but is extremely rare; 'D' indicates the attested absence of the feature; 'X' means that the feature is not applicable given the structural make-up of the variety; finally, '?' stands for missing information. Since this type of representation is typical of ordinal data, the letters were transformed to an ordinal scaled categorical variable with 'X' and '?' recoded as 'NA' (Levshina 2015: 343). It needs to be noted from the start that the only Canadian variety found in the atlas is Newfoundland English which cannot be taken as representative of the whole of Canada (see Boberg 2008: 146) and was subsequently left out of the comparison. Also, BrE is not represented as one variety but rather as several regional dialects in the atlas data. For the purpose of the present comparison, English spoken in Southeast England was used as a proxy (Anderwald 2004: 175). All other varieties have direct equivalents in the atlas data. The distance between these eight varieties (without Canadian English) was calculated with the Gower general coefficient of similarity using the `daisy()` function in the cluster package (Maechler et al. 2016) in order to take the ordinal scale of the data into account (Levshina 2015: 343). Dimension reduction was computed with `isoMDS()` from the MASS package (Venables & Ripley 2002).

The reduction of the 8×8 dimensional distance matrix resulted in a two-dimensional MDS map that clusters PhiE and JamE close together, NZE and BrE close together, IndE, SinE and HKE together and plots IrE separately from the rest (see Figure 5.32). This pattern does not correspond to any of the previously observed clusters. This visually low correlation with earlier MDS solutions finds empirical support in mantel tests: The correlation with the 1st line (significance) is moderate ($r = 0.5314$) and significant ($p = .024$); the correlation with the 2nd line (effect sizes) is minimal ($r = 0.1936$) and not significant ($p = .279$); the correlation with the 3rd line (constraint hierarchy) is also minimal ($r = 0.1047$) and not significant ($p = .35$). (CanE was excluded from the tests.) Since the distances obtained on the basis of coefficient estimates are the only measure that correlate significantly with the distances derived on the basis of morphosyntactic features, and because the bootstrapping also indicated coefficient estimates to be the most reliable measure of comparison, it is likely that the 2nd line of evidence constitutes the best measure to assess changes in probabilistic grammars that are linked to overt structural changes.

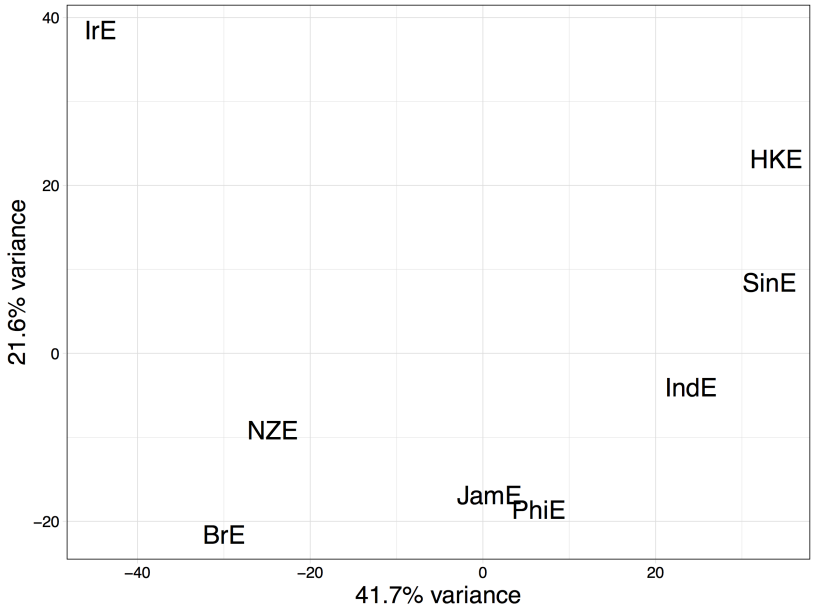


Figure 5.32 Multidimensional scaling map of eight varieties of English, based on the morphosyntactic feature catalogue accompanying the *Electronic World Atlas of Varieties of English* (Kortmann & Lunkenheimer 2013) — Distances between varieties correspond to the aggregate morphosyntactic dissimilarity between varieties.

The next step now would be to cluster these probabilistic grammars for each of the three dimensions in order to obtain a measure of how many probabilistic grammars can be distinguished. This, however, is beyond the scope of the current study and remains to be addressed in future work.

5.7.5 Interim summary

To conclude, instead of finding a stable probabilistic grammar across varieties, we have mainly observed divergence on three different levels of comparison and a surprisingly small degree of global similarity. Especially with regard to effect sizes, average similarity was lower than 0.6 while for the other two dimensions of comparison, statistical significance and constraint hierarchy, the average similarity score was 0.756 and 0.839 respectively. In contrast to, for instance, the genitive alternation, the probabilistic grammar underlying the dative alternation seems much more fluid in a regional comparison and on the three levels of comparison.[1] No lectal clusters

[1] As follow-up studies show, the stability scores very much depend on the predictors included in the mixed-effects models and random forests.

emerged on the basis of varieties' shared socio-historical background (for instance clusters of Asian varieties), nor could we observe that the distances from one analysis coincided with the distances from the other analyses. Instead, all three comparisons distinguished varieties with North American roots from the rest of the varieties. Finally, the distance measures obtained from the three comparisons also did not coincide with the distance measures obtained from the catalogue of 76 morphosyntactic features with the exception of effect sizes. In light of these findings, the extent to which we can even talk about a stable probabilistic grammar is questionable. Rather, we seem to observe several factors that impact linguistic variation on various levels of speakers' probabilistic grammar. This variability clearly merits further elaboration and will thus be addressed in the discussion in the next chapter.

5.8 Chapter summary

The current chapter aimed to more closely probe regional variation in the probabilistic constraints impacting the choice of dative variant. To that end, random forests and mixed-effects models were fitted to assess the relative importance of predictors and their cross-regional malleability in detail. The conditional random forest fitted to the dataset revealed length and recipient pronominality to be the most important predictors; it is also length and recipient pronominality (as well as CORPUS) that turned out to be regionally variable in their effect size. All constraints in the mixed-effect model (including length and recipient pronominality) have a congruent effect on dative choice: Speakers consistently prefer that dative variant in which the first constituent is easier to produce than the second one – easier meaning animate, definite, pronominal, short(er), etc. While the effect **direction** is thus constant cross-regionally, the effect **size** of three constraints, namely WEIGHTRATIO, RECPRON and CORPUS, differs from the global average in four varieties, that is, in Indian, Hong Kong, Irish and Jamaican English, to varying degrees (summarised for convenience again in Table 5.29).

The malleability of weight ratio, recipient pronominality and CORPUS was further explored with a closer investigation of these three predictors. First, the coding of end-weight was amplified with a fine-grained complexity coding which was observed to impact dative choice independently of length. The results of that analysis further revealed that not all weight-related measures are regionally malleable, that is, length is regionally malleable but complexity is not. The second detailed analysis zoomed

Table 5.29 Summary of cross-varietal differences in effect size — Minus (-) indicates decreased effect size, plus (+) indicates increased effect size.

Variety	WEIGHTRATIO	RECPRON	CORPUS
IrE	–	=	=
IndE	–	+	–
JamE	+	=	=
HKE	=	–	=

in on the effect of recipient pronominality across varieties and offered comprehensive insights into the variability in lexical profiles of dative variants across varieties. Covarying and distinctive collexeme analyses showed that the average collexeme strength between the lexical items in the dative variants is always higher in the prepositional dative compared to the ditransitive dative. While no cross-regional differences emerge in the covarying collexeme analysis, the distinctive collexeme analysis highlights a non-native versus native variety split with regard to pronominal recipients: Pronominal recipients – especially *you* – are more strongly attracted to the ditransitive dative in non-native varieties compared to native varieties. A similar native versus non-native split in the ditransitive dative was observed with respect to the mean collostructional strength of themes. In contrast, regional stability largely prevails in the association strengths of verbs. These findings suggest that speakers of non-native varieties seem to follow the principle of recycling lexical items in the ditransitive dative, especially with regard to the recipient. In other words, ditransitive datives are more concrete and lexically predefined constructions in non-native varieties compared to native varieties. The third analysis took the cross-varietal differences in the effect of CORPUS as a starting point to probe the register-specificity of the English dative alternation further. Results thereby indicate that register influences syntactic variation in the dative alternation differently across varieties. Most clearly, the prepositional dative is more likely in a majority of registers in non-native varieties (HKE, IndE, JamE, PhiE, SinE) while register effects in native varieties decrease the likelihood of a prepositional dative below the 50% threshold. Additionally, results of that third step show that probabilistic constraints, first and foremost recipient pronominality, are not only regionally but also stylistically malleable and thus turn out to be truly cross-lectally variable. Finally, on the basis of the observed variability in probabilistic grammars, the differences and similarities in varieties' probabilistic

grammar were assessed quantitatively. For that purpose, I applied methods from comparative sociolinguistics to quantify the probabilistic distance between varieties and to calculate a stability score that is indicative of the extent to which we can find a shared probabilistic grammar across regionally distinct varieties of English. Probabilistic distances were calculated along three dimensions that measured different aspects of probabilistic grammars, namely statistical significance, effect size and relative importance of predictors. Instead of finding a stable probabilistic grammar across varieties, the analysis highlighted divergence on the three levels of comparison and a surprisingly small degree of global similarity. Especially with regard to effect sizes, average similarity was lower than 0.6 while for the other two dimensions of comparison, statistical significance and constraint hierarchy, the average similarity score was 0.756 and 0.839 respectively. In contrast to, for instance, the genitive alternation, the probabilistic grammar underlying the dative alternation seems more fluid in a regional comparison (average stability score of 0.729 in the dative versus 0.829 in the genitive alternation). Apart from a cluster of North American (influenced) varieties (CanE and PhiE), no lectal clusters emerged on the basis of varieties' shared socio-historical background, nor could we observe that the distances from one analysis coincided with the distances from the other analyses. What is more, the distance measures gained from the three comparisons did not coincide with the distance measures obtained from the feature catalogue of 76 morphosyntactic features accompanying the *World Atlas of Varieties of English* (Kortmann & Lunkenheimer 2013) apart from a significant but only moderate correlation with the distances obtained from the comparison of coefficient estimates. In the light of these findings, the extent to which we can thus even talk about a stable probabilistic grammar is questionable. Rather, we seem to observe an entanglement of factors that impact linguistic variation on various levels of speakers' probabilistic grammar. The results presented here have thus shown that while stability prevails on the macro-level concerning the effect direction and mainly also the size of constraints, even subtler differences emerge with regard to levels of significance, effect sizes and ranking of predictors.

6

Discussion[1]

This chapter first offers a summary of the current study from the introduction to the final results before moving on to address and answer the research questions posed at the beginning of the study. Results that point to broad-ranging patterns in regional variation will take centre stage. Findings are then discussed within the broader context of language production and comprehension and three tentative explanations for patterns of variation are suggested. The findings are next grounded in previous corpus-based work on the English dative alternation and differences and similarities with the results of those earlier studies are reviewed. The chapter finishes with an overview of the innovative aspects – including the study's contribution on the descriptive, methodological and theoretical plane – as well as with a discussion of the limitations of the present work.

6.1 Summary

This study has set out to explore the cross-lectal malleability of the underlying probabilistic constraints that shape variation in the dative alternation, that is, the variation between the ditransitive dative (e.g. *John gives Mary the apple*) and the prepositional (e.g. *John gives the apple to Mary*), in nine national varieties of English from a probabilistic and functional-cognitive perspective.

To that end, possibly alternating dative variants were extracted from nine compo-

[1] Parts of this chapter are based on Röthlisberger et al. (2017)

nents of the International Corpus of English (ICE) and the Corpus of Global web-based English (GloWbE) using a verb list of 86 alternating verbs, thus tapping into patterns of variation in the dative alternation in British English, Canadian English, Hong Kong English, Indian English, Irish English, Jamaican English, New Zealand, Philippine English and Singapore English. Following standard practices in Variationist Sociolinguistics (Tagliamonte 2012), the dataset was then restricted to those dative observations that did occur in the envelope of variation, that is, in a context whereby the alternating variant would be grammatically acceptable and semantically equivalent, leaving 13,171 dative tokens to be coded. After defining the boundaries of the verb and the two objects, the heads of the noun phrases were identified and each dative variant was annotated for language-external as well as language-internal factors. The language-external factors comprise information provided by the corpus structure of ICE on register, genre, file number, speaker number (within the file), text number and mode (spoken vs. written). The GloWbE data was integrated into that categorisation as online written data. Further included were the predictor CORPUS to distinguish between ICE and GloWbE, and VARIETY, that is, the nine varieties from which the data was sampled. Language-internal factors were mainly coded following previous literature (see, for instance, Bresnan et al. 2007a; Wolk et al. 2013). Recipients and themes were semi-automatically coded for animacy, definiteness, information status, noun phrase expression type and their length in the number of letters. The whole construction was additionally for VERBSENSE. Due to sparseness of data in some predictors, the number of levels had to be reduced. As such, the five-level distinction in animacy was merged to two ('animate' vs. 'inanimate') and the six-level predictor that gauged noun phrase expression type was reduced to the binary predictor to gauge pronominality, namely RECPRON and THEMEPRON. Extending previous accounts of the dative alternation, the dative tokens were further annotated for the syntactic complexity of recipient and theme, syntactic priming (persistence), lexical density (type-token ratio) of the variant's context, lexical frequency of the recipient and theme in general and thematicity (frequency in the specific text) of recipient and theme. Besides these more abstract annotations, the lexical items (head of recipient and theme, verb lemma) were also coded separately in the data frame to be included as random effects in the logistic regression.

The study made use of six investigative strands to assess the multifaceted nature of the constraints that shape variation in the dative alternation, the results of which were presented in separate sections in the preceding chapter.

In the first step, conditional random forests were fitted to the data (as a whole but also by variety) to establish the ranking of predictor importance. That analysis found relative length and recipient pronominality to be the most prominent factors with length carrying more weight than recipient pronominality in all varieties except Indian English where the order in prominence was reversed.

In a second step, a mixed-effects model was fitted to the data using the predictors summarised above in order to explore the extent to which language-internal factors differ significantly in their effect sizes or directions across the nine varieties. The results of that analysis highlight the overall prevailing regional stability in the factors driving the variation between the prepositional and the ditransitive dative variant, thus corroborating results of previous corpus-based accounts of the dative alternation. Results also indicate that the effect of the predictors in the study behave generally as expected given the literature in that speakers tend to favour that dative variant where the first constituent is given, pronominal, short, animate, definite and simple and the second constituent is new, nominal, longer, inanimate, indefinite and complex (see Bresnan et al. 2007a). At the same time, three factors turned out to be cross-regionally malleable in their effect size, namely relative length, recipient pronominality and corpus. The nature of these constraints' variability was investigated separately in the subsequent three steps.

Since relative length of constituents only served as a proxy for end-weight effects, the operationalisation of end-weight was evaluated and extended by a second measure – structural complexity – to complement length measurements. The coding of structural complexity (NP-structure) followed Berlage (2014) and included a fine-grained five-level distinction that took both the *nouniness* of the constituent as well as the number of post-head dependents into account. Mixed-effects models and random forests were fitted to compare the effect of structural complexity and relative length on dative choice and to assess the extent to which structural complexity is regionally malleable. The two analyses revealed that length is a better predictor than structural complexity on a global as well as local level, and that structural complexity (as operationalised here) is not regionally malleable.

Next, the variability of recipient pronominality was taken as a starting point to assess the degree to which lexical constraints on the dative alternation varied regionally. To that end, two techniques from the family of collostructional analyses were used: Covarying collexeme analyses measured the strength of association between two lexical items within a specific dative variant; distinctive collexeme analyses gauged

the strength of association between the lexical items and a specific dative variant. The latter method in particular was beneficial to probe the cross-regional variability of the effect of recipient pronominality observed in the first mixed-effects model. Findings from the covarying collexeme analysis show that mutual attraction between lexical items in the dative alternation (between verbs, recipients, themes) is on average higher in the prepositional dative than in the ditransitive dative with collexemes in Hong Kong and Indian English showing the highest collocational strength in the prepositional dative. Findings from the distinctive collexeme analysis indicate that verbs and recipients are closely associated with the ditransitive dative and themes more closely with the prepositional dative (on average). *Give* turns out to be the prototypical ditransitive verb in line with previous research, and *pay* is the verb most closely associated with the prepositional dative. While marginal regional differences emerge in theme and verb collostructions, it is with the recipient that the most pronounced differences can be observed: The list of recipients most closely associated with the ditransitive dative is headed by seven pronominal recipients. Association strength is highest in non-native varieties and lower in native varieties where recipients are more closely associated with the prepositional than the ditransitive dative on average. Overall, the results suggest that the ditransitive dative is lexically much more entrenched for speakers of non-native varieties and that this entrenchment depends on both the recipient and theme.

In the penultimate step, the statistically significant difference between ICE-India and GloWbE-India regarding the choice of dative variant provided the grounds to zoom in on the register-specificity of the English dative alternation. The main focus was thus on the effect of register on dative choice and the extent of cross-register variation in the effect of language-internal constraints. Two separate mixed-effects models were fitted to first gauge the effect of register on dative choice and second, to explore the intra-systemic variability of probabilistic grammars across different registers. REGISTER distinguishes between the four registers sampled in ICE ('printed', 'non-printed', 'dialogue', 'monologue') and the one register provided by GloWbE ('online'). The results of the first model indicate that register impacts dative choice significantly in Hong Kong, Indian, Irish and Jamaican English. What is more, the majority of registers in non-native varieties increase the likelihood of a prepositional dative in contrast to native varieties (where the ditransitive is more likely). The second model shows that language-internal constraints are also stylistically variable; first and foremost recipient pronominality displays instability in its effect size across

the five registers, highlighting the cross-lectal malleability of this constraint.

Finally, the gradience of probabilistic grammars was probed in a sixth and final step to assess the stability of probabilistic grammars using three proposed measures from Comparative Sociolinguistics (Tagliamonte 2002). The first measure compared the nine varieties based on the number of shared significant and non-significant predictors. For that purpose, nine mixed-effects models were fitted separately per variety that included the five most important predictors given by the random forest fitted on the full dataset. The distance between varieties was then calculated using Manhattan distance and visualised with multidimensional scaling techniques. No regional clusters (or native versus non-native varieties) emerged from that comparison apart from a slight indication of a North American cluster involving PhiE and CanE. The overall stability, that is, similarity score between varieties amounted to 0.756. The second measure compared the varieties based on the effect size obtained from the same nine mixed-effects models used in the first comparison. Distances between varieties were calculated using the Euclidean distance metric and visualised with multidimensional scaling. This time, variety clusters seemed more clear-cut: CanE and PhiE form a tight cluster of North American (influenced) varieties and JamE is – as the only creole-based variety – plotted away from the rest. Average stability score amounted to 0.591. The third and last comparison used the relative importance of predictors, that is, the constraint ranking, of by-variety random forests fitted with the same set of five predictors. Distances between varieties' rankings were calculated with spearman's ρ and visualised with multidimensional scaling. In this third comparison, the North American cluster became visible again. IndE was plotted away from the rest of the varieties. Stability score amounted to 0.839. The average stability score across all three lines of comparison finally added up to 0.729 which indicates some stability (value above 0.5) but also more variability than, for instance, in the genitive alternation (average stability score of 0.829, see Heller 2018). Finally, a comparison between the probabilistic distances calculated on the basis of statistical significance, effect sizes and constraint ranking revealed hardly any correlation with traditionally calculated linguistic distances based on morphosyntactic features.

In sum, two patterns of note have been uncovered in the current study: Probabilistic grammars are stable across nine geographically diverse varieties of English with respect to the effect direction of constraints. On the other hand, we observed small differences in the degree of sensitivity that speakers of different varieties demonstrate towards some of the factors that constrain variation as well as in the significance,

effect size and constraint rankings underlying probabilistic grammars. Results have furthermore shown that these factors operate independently of correlated predictors (in the case of length and complexity) and can be tied to lexical effects (in the case of recipient pronominality). Concerning the gradience that this study found in probabilistic constraints, no exhaustive explanation can be provided at this point for the regional variation that was observed. Rather, three somewhat speculative but plausible explanations are suggested in the subsequent section on how such variation might arise as the result of random but expected modifications in the effect sizes of predictors through the constant reuse of structural patterns (such as collostructions) that speakers are exposed to.

6.2 The gradience of probabilistic grammars — three suggestions

The results of this study have emphasised the pervasiveness of variation in probabilistic grammars across different levels of comparison while also observing stability in the effect direction of constraints. Thus coming back to the first research question, that is, the extent to which we find a shared probabilistic grammar, the study reveals that lectal variation is ubiquitous in language even in subtle stochastic constraints that shape linguistic variation. The study has also shown that, concerning probabilistic constraints, no clear-cut difference can be made between varieties on socio-historical or evolutionary grounds. Rather, all varieties display variation in probabilistic grammars to various degrees and at various levels of subtleties: The results from the mixed-effects model distinguish CanE and IrE versus HKE and IndE regarding constructional preferences and foreground differences in effect sizes for IndE, IrE, HKE and JamE. Typologically robust predictors, such as length for instance, turn out to be cross-lectally malleable, not just across varieties but also across different registers (third research question). It is in the lexical profiles of dative variants where more extensive cross-regional variation is discernible. Here, more so than in the probabilistic domain, can we observe a split between native and non-native varieties of English. Collexeme and collostructional analyses set apart HKE, IndE, NZE and CanE with respect to collocational strengths, HKE and IrE regarding the strong association of verbs with ditransitive datives, HKE and SinE regarding the strong association of *give* with the ditransitive dative and HKE and JamE regarding the strong association of *pay* with the prepositional dative. Furthermore, the collostructional analysis of

recipient associations highlighted the strong relation between pronominal recipients and the ditransitive dative in non-native varieties in contrast to native varieties. A similar split between native and non-native varieties was also found in the register analysis. The analysis of register effects foregrounds HKE, IndE, IrE and JamE – the only four varieties where register has a significant impact on dative choice – and BrE, IndE, NZE and PhiE with respect to the malleability of recipient pronominality across registers. The cross-lectal malleability of recipient pronominality thus addresses the fourth research question, namely which of the individual constraints are tied to stylistic differences or lexical considerations. As shown by the present study, recipient pronominality turned out to be variable across styles in that its effect pervasively deviated across various registers in four varieties of English. The effect of recipient pronominality is also tied to lexical considerations, as shown in Section 5.5 on collostructions. Finally, when looking for a measure of stability of probabilistic grammars, the probabilistic distances calculated from the comparison of statistical significance, effect size and constraint ranking often clustered CanE and PhiE closely together. In sum, only with regard to lexical effects and register effects do native versus non-native differences become most apparent, raising the question whether lexical effects and stylistic differences are decisive factors in a native vs. non-native split due to their structural overtness and prominence. Structural innovations involving lexical items are the most palpable outcomes in new emerging varieties of English. And while stylistic differences are said to be minor across varieties, especially when comparing the more formal registers (see Hundt et al. 2016), notable fine-grained differences have also been discerned (see, for instance, Ehret 2008). Hence, while we might observe effects of fluidity in probabilistic grammars in various varieties for different reasons, structural preferences and stylistic differences are probably the most effective to separate non-native from native varieties. Other analyses presented here have not found such a clear split. Instead, by digging deeper into probabilistic grammars the striking ubiquity of lectal variation becomes apparent. So, how much deeper do we have to dig to find stability and where does lectal variation stop? In view of the current findings, the limits on cross-varietal variation might well be situated in the individual or social groups within a speech community rather than on the level of the speech community as a whole. While the present analysis has unraveled inter- and intra-dialectal variation (between regions and within a region), inter-individual and intra-individual lectal variation remains a mystery which the current dataset, due to the sparsity of data from individual speakers, cannot attempt to address. Probing the

lower boundaries of lectal variation in probabilistic grammars thus forms a desirable asset of future work.

This leaves us with the second research question, namely whether this unsystematic variability is random or, if not, whether it can be explained by considering socio(-historical) factors. If, as probabilistic approaches fundamentally argue, grammatical knowledge includes a probabilistic component, and if, as usage-based approaches to language argue, language and therefore grammatical knowledge is acquired from language experience, two predictions follow: First, stochastic regularities are derived from language production and comprehension, and second, subtle shifts in speakers' linguistic experience can lead to gradient yet detectable variation in these underlying stochastic regularities. These two predictions can be verified in more detail by drawing on general biases in language production and planning.

According to MacDonald (2013), incremental language production can be explained by the interplay between three principles: *Easy First, Plan Reuse* and *Reduce Interference*. An *Easy First* bias in speech production and planning leads a speaker to select early those linguistic units (words, phrases and so on) that are easier to retrieve from long-term memory. 'Easier' in this sense is typically characterised as frequent, shorter, less syntactically complex, conceptually entrenched and given in the discourse (MacDonald 2013: 3). At the same time, speakers tend to reuse previously heard syntactic plans and closely related structures that they retrieve from long-term memory in a process that MacDonald (2013: 4) calls *Plan Reuse*. The third process, *Reduce Interference*, refers to the minimisation of interference from a semantically closely related lexeme during the utterance of a word by increasing the number of linguistic units between the two words. MacDonald argues that these three principles of language production and planning jointly govern utterance form. For instance, animate nouns have been shown to be easier to retrieve from memory than inanimate nouns, hence the tendency for animate agents to be realised in subject position, as in *The boy smashed the window* (Bock 1982). At the same time, passive sentences, as in *The window was smashed by the boy*, often involve inanimate subjects since patient arguments tend to be inanimate. As the forces of Easy First (animate first) and Plan Reuse (priming for passive voice with a passive biased verb) might conflict in the choice of passive vs. active, we expect utterance planning time to increase for passive voice – a prediction supported by experimental evidence (Ferreira 1994). These three principles do not only jointly constrain utterance form but (over time) also generate the link from individual-level behaviours to population-level linguistic phenomena

(Scott-Phillips & Kirby 2010: 411). By summing over millions of utterances and language producers, the consistent interplay between the three principles creates statistical regularities in language usage (MacDonald 2013: 5).

The outcome of this interplay is reflected in the statistical models. On the one hand, speakers tend to choose the dative variant in which the first constituent fulfils all aforementioned requirements of being 'easy'. Easy First is thus a principle that combines the various influences of the factors in our model, such as length, frequency, givenness and definiteness. Since the combination of these factors constitutes language users' probabilistic grammar, we can assume that the prevailing stability in effect direction that we observe across regional varieties of English can be attributed to the principle of Easy First. On the other hand, while Easy First seems to strengthen stability of speakers' probabilistic grammar irrespective of the linguistic material, Plan Reuse constantly reinforces the regularisation of linguistic input. However, if this linguistic input varies between different lects, Plan Reuse will strengthen diverging statistical patterns of use. Changes in the linguistic material can thus result in differences in the statistical regularities that speakers make and eventually in diverging probabilistic grammars. As a consequence of these diverging statistical regularities, the strength of the effects of the individual predictors that modulate these regularities change as well. Hence, which (syntactic) variant is cued and thus easier for speakers to produce or entrenched enough to be reused in language planning may not necessarily be the same for speakers but will depend on their individual linguistic experience (Ellis 2002: 145). At this point, I cannot profess to be able to provide exhaustive explanations for the regional variation in the strength of some predictors that we have observed. Rather, I would like to suggest three somewhat speculative but plausible explanations of how such variation might arise as the result of (random but expected) modifications in the cue strength of predictors through the constant reinforcement of structural patterns by Plan Reuse.

First, linguistic experience and input vary due to the general conditions of language or dialect contact, which naturally vary from region to region as speakers of different dialects and/or native languages interact in their new environment. Such contact leads to the emergence of localised linguistic forms on the level of syntax and morphology in the formative stages of New Englishes – a process that Schneider (2007: 44) calls "structural nativisation". Structural nativisation generally results in new combinations of syntactic constructions with lexical items. In cases where new lexical items occur frequently enough in these syntactic constructions, the abstractions of regularities

that speakers make (in order to be able to generalise beyond the linguistic input) lead to changes in the constraints governing language structure. These constraints are, in turn, learnt during processes of language acquisition (Ellis 2002: 144) and become part of speakers' grammatical knowledge (Gahl & Garnsey 2004). In short, changes in lexical choices in syntactic variants can influence the impact that the underlying cues have on syntactic variation.

Second, processes involved in second language acquisition and substrate influences may also shape users' choices in a given context. Note that some of the largest deviations in individual factor effects in the model occur in the L2 varieties; it is in IndE, HKE and JamE (and IrE) where the effects of weight ratio and recipient pronominality deviate significantly from the global average. Effects of second language acquisition impact not only structural nativisation processes but also lead to an increased usage of the more transparent syntactic variant – in our case the prepositional dative (Leufkens 2013: 345-346; see also Siegel et al. 2014). This in turn can lead to changes in the strength of specific cues as variants are used by L2 speakers in contexts where L1 speakers would not use them. For instance, Mukherjee & Hoffmann (2006) explain the large proportion of prepositional datives in IndE by drawing attention to the fact that *give* frequently occurs as a light verb in that variety, as in (75). They also show that the kind of verb-complementation profiles that *give* is used with in IndE differs from British English (Mukherjee & Hoffmann 2006: 154-155). De Cuypere & Verbeke (2013: 180-181) further suggest that the popularity of light verbs in IndE is due to their high frequency in the substrate languages. In addition, the necessity of an explicit dative case marker in the Indian vernacular languages (for instance, *ko* in Hindi as in 76) might have increased the use of the prepositional dative in IndE in contexts diverging from L1 usage (see also Haspelmath 2013).

(75) *give a satisfactory and convincing explanation to any one of them* <ICE-IND:W1B-016>

(76) Hindi
*maiṃ apnī bahan-**ko** yah kitāb deti hūṃ.*
I my sister=to.REC the book.TH give.
'I give my sister the book.' (De Cuypere & Verbeke 2013)

A similar substrate effect can be observed in the contact situation between Jamaican Creole and Jamaican English. According to the *Atlas of Pidgin and Creole Language*

Structure (APiCS) Online (Michaelis et al. 2013), speakers of Jamaican Creole use ditransitive constructions as in (77) with verbs of physical transfer of possession followed by recipient and theme without any additional grammatical marking on the recipient (contrary to what one would expect in Standard English).

(77) Jamaican Creole
Di uman gi di bwai di fuud.
DET woman give DET boy.RECIPIENT DET food.THEME.
'The woman gave the boy the food.' (Farquharson 2013)

Bruyn et al. (1999: 330) provide examples from several other creoles that highlight that the ditransitive variant with an unmarked recipient constitutes the most frequent if not only option in creole languages, irrespective of whether the recipient occurs before or after the theme. The high frequency of ditransitive variants seems to be inherent to creoles, independent of the fact that not all lexifier languages had those ditransitive variants to begin with. Since most speakers in India and Jamaica acquire the substrate language as their first language (see Meade 2001: 175-176 for the Jamaican context), transfer effects would result in ditransitive and prepositional datives being used in different contexts in both IndE and JamE. In addition to the transfer of structural preferences from one's native language, transfer of cue strength (that is, the effect size of constraints) from the first language can also lead to gradient shifts in linguistic preferences and changes in speakers' probabilistic grammar (MacWhinney 1997: 129).

Third, the variation we observe might not only be due to changes in contact-induced lexical variation or substrate effects but also result from constructional and/or semantic changes that arise in the course of everyday language usage. As speakers use the ditransitive or prepositional datives in different ways in different contexts, the range of meanings associated with either variant – their semasiological profiles – will likely change, and these changes are reflected in the lexical items that fill their syntactic slots. This entails that the range of different lexical items might be more diverse in one variant compared to the other and that this difference in diversity (that is, the degree of semasiological heterogeneity) might differ from variety to variety. The latter hypothesis is supported by studies that show that universal processes of language acquisition can influence the type frequency in syntactic variants. For instance, research in first language acquisition has shown that up to a certain age, children associate the use of the ditransitive dative with specific lexical items and do

not abstract to other syntactic constructions beyond the input they receive (Dodson & Tomasello 1998: 606). Similarly, second language learners tend to associate the use of the ditransitive dative with specific lexical items (for instance, pronouns) or certain discourse contexts while the use of the prepositional dative is not as semantically restricted (McDonough 2006: 193-194). The findings of the distinctive collexeme analysis, that showed that non-native speakers of English tend to lexically entrench the ditransitive dative, certainly attest to that as well.

In the end, the constant reinforcement of such diverging usage patterns through the principle of Plan Reuse can result in diverging statistical regularities. This process has been termed *probabilistic indigenisation* by Szmrecsanyi and colleagues (2016) drawing on the concept of structural nativisation established in the World Englishes paradigm (Schneider 2007). Szmrecsanyi et al. (2016: 133) define probabilistic indigenisation as

> the process whereby stochastic patterns of internal linguistic variation are reshaped by shifting usage frequencies in speakers of post-colonial varieties. To the extent that patterns of variation in a new variety A, e.g. the probability of item x in context y, can be shown to differ from those of the mother variety, we can say that the new pattern represents a novel, if gradient, development in the grammar of A. These patterns need not be consistent or stable (especially in the early stages of nativization), but they nonetheless reflect the emergence of a unique, region-specific grammar. (Szmrecsanyi et al. 2016: 133)

Röthlisberger et al. (2017) also draw on that definition but stress the outcome of probabilistic indigenisation, namely the emergence of a unique, region-specific grammar. They refer to this emergence as *cognitive indigenisation* to refer to the lectalisation or creation of distinct lects at the level of very subtle gradience. Cognitive indigenisation is thereby closely connected to and even dependent on shifting usage frequencies in the language variety. For instance, in IndE, we observed that recipient pronominality is a very strong cue for the choice of dative variant and plays a crucial role in the probabilistic indigenisation process in that specific variety. The reason for the strong cue validity of recipient pronominality on the choice of dative variant in IndE is reflected in the fact that speakers of IndE are exposed to a large number of ditransitives with a pronominal recipient. Cross-varietal differences with regard to the variants' lexical profile can thus lead to deviations in the underlying factors that constrain linguistic variation.

That the operation of linguistic constraints is limited by lexical considerations

is nothing new (see Bybee & Hopper 2001: 2). I have shown, however, that the strength of these lexical constraints varies subtly between different varieties of the same language. Cross-regional variation in the lexical effects finds support in the wide range of by-random effect adjustments of the by-variety mixed effects models fitted for the probabilistic stability scores. These mixed-effects models include the five most important predictors given the results of the random forest in Section 5.2 and random intercepts for file number, verb, recipient and theme heads. Plotting the by-random effect adjustments to the intercept for each of the nine models makes the lexical variability noticeable (see Figure 6.1). Variation in the lexical constraints on dative choice is most extensive regarding the random adjustments by verb (bottom figure in Figure 6.1) and less so for the random adjustments by theme (middle figure in Figure 6.1). Recipients hardly contribute at all to the model fit (top figure in Figure 6.1). Also note that Irish English exhibits the largest adjustments for both verbs and themes.

Recipient pronominality and length are not only the two factors that differ significantly across varieties (and registers), they are also the most influential constraints on dative choice on a global scale. The findings of the present study thus suggest that the factors that emerge as the most amenable to probabilistic indigenisation are also the most prominent cues, namely those factors that carry the most cue validity. Hence, even though we might never be able to fully predict which factors in linguistic variation might deviate across different dialects or varieties, we can assume that the most reliable cues are the ones most probably prone to change in strength (see similarly Heller 2018 and Grafmiller 2014). Why is it that HKE, IndE, IrE and JamE exhibit the greatest difference? While the forces of structural change suggested here might point us into the direction of the reasons for different degrees of probabilistic indigenisation across varieties, I cannot conclusively answer that question with the data currently at hand. Furthermore, cross-constructional comparison reveals that the set of varieties that diverge the most from the global mean is not consistent (see Szmrecsanyi et al. 2017). If – as suggested above – the lexical profile of the variants influences the statistical abstractions that speakers make, and assuming that the lexical profiles differ from variant to variant and from alternation to alternation, we can expect construction-specific statistical deviations in the influence of different predictors across varieties. Since both ICE and GloWbE do not sample the linguistic system of a variety as a whole, it is not surprising then that my findings deviate to some extent from previous work.

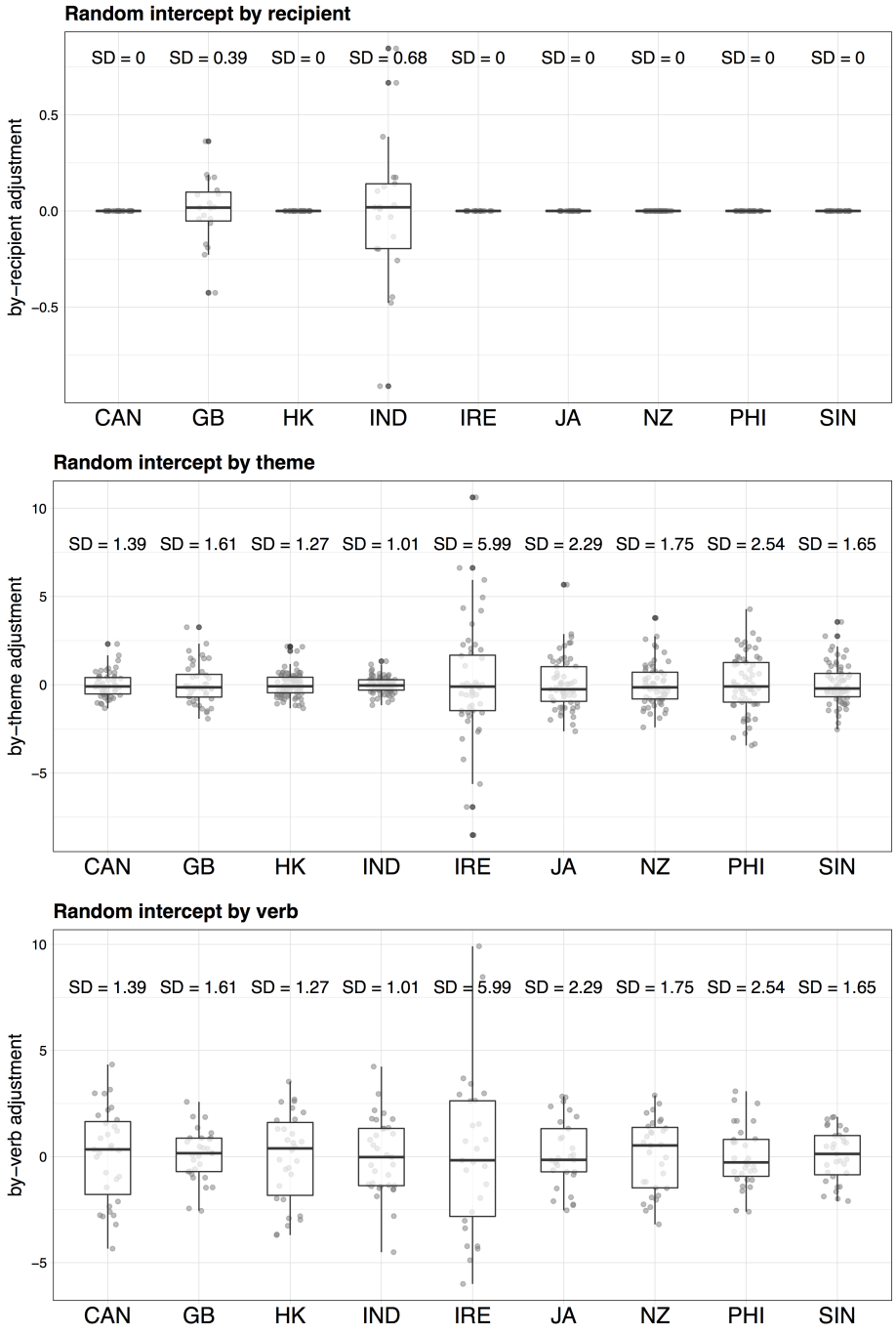


Figure 6.1 (a) top: By-recipient adjustments to the intercept; **(b) middle**: By-theme adjustments to the intercept; **(c) bottom**: By-verb adjustments to the intercept — Varieties are ordered alphabetically.

6.3 Reflecting and extending previous research

The analyses undertaken in Chapter 5 provided insights into various aspects of lectal variation on the probabilistic level of the language system. The majority of these aspects had already been touched upon in earlier works on the dative alternation.

One of the main aspects that this study investigated was the extent to which probabilistic constraints fuelling variation between the two dative variants are subject to cross-lectal (regional as well as stylistic) variability. In that regard, the findings of the current study largely substantiate results of previous corpus-based analyses of the dative alternation. Similar to previous work, my study highlights a prevailing cross-varietal stability in probabilistic grammars in the effect direction of constraints (e.g. Schilk et al. 2013; Bernaisch et al. 2014). What is more, the cross-lectal variability of the effect of length and recipient pronominality presented here were also already observed in earlier work (Schilk et al. 2013: 14).

The present study largely confirms relative length as the locus of probabilistic differences: Wolk et al. (2013) record a significant difference in the effect of theme length between British and American English, an effect which the current study finds for Indian, Irish and Jamaican English. One has to keep in mind, however, that Wolk et al. (2013) investigate historical written corpus-based data while the present study analyses contemporary spoken and written data. The regional variability of length effects is also confirmed by Szmrecsanyi et al. (2017) who report eight probabilistic contrasts as a result of pairwise regression comparisons between a total of four varieties of English. They observe significant differences in length effects in three pairwise comparisons, namely American English vs. Canadian English, British English vs. Canadian English and Canadian English vs. New Zealand English (regional contrasts that are all not confirmed by the current study). In addition to length, Szmrecsanyi et al. (2017) observe a significant difference between AmE and BrE and between BrE and CanE with regard to the effect size of recipient pronominality, and between AmE and NZE, between AmE and CanE and between BrE and NZE regarding the semantics of the verb. While the present study confirms the regional variability of length effects and recipient pronominality observed in earlier work (see Schilk et al. 2013: 22), the study could not attest the exact same probabilistic constraints contrasting significantly between the same set of varieties as reported in Szmrecsanyi et al. (2017) or other research. Besides verb semantics (Szmrecsanyi et al. 2017), one such contrasting probabilistic constraint is recipient animacy which turned out to have a significantly

different effect in New Zealand English vs. American English (Bresnan & Hay 2008) and in Canadian English vs. British English (Tagliamonte 2014). Bresnan & Hay (2008) report that speakers of New Zealand English are more sensitive to recipient animacy than speakers of American English. Tagliamonte (2014) observes a similar difference in effect size between speakers of Canadian and British English with a weaker effect of recipient animacy in Canadian English.

The deviations of the current study's results from previous findings is not surprising since the majority of earlier work focuses on one or two varieties (e.g. Bresnan & Hay 2008; Bresnan & Ford 2010; De Cuypere & Verbeke 2013), often samples from one specific text register (e.g. Schilk et al. 2013; Szmrecsanyi et al. 2017) or restricts attention to the prototypical verb *give* (e.g. Bernaisch et al. 2014). And of course, differences in coding practices might additionally confound the results. Two suggestions follow from this: First, in order to fully understand the cross-lectal plasticity of the probabilistic factors shaping variation in dative grammar(s), an aggregate perspective as the one adopted in this study is necessary. While I admit that the data sources subject to study here are not as basilectal as the data in, for instance, Szmrecsanyi et al. (2017), the present study nevertheless shows that pooling over a large number of verbs and varieties offers a more aggregate perspective and generalisability of the results. Second, potential meaningful differences between varieties could remain hidden with such an aggregate measure if we do not zoom in on the various detailed aspects of variation. Not only are some effects sensitive to the lexical items that are used as syntactic constituents, some effects also seem to be specific to certain registers. Multiple fine-grained studies would thus be needed to look at all aspects of the variation. Needless to say that the dataset used for the present aggregate perspective offers the possibility for such fine-grained studies.

The lexical sensitivity of probabilistic constraints is not only relevant with respect to the stronger association of pronominal recipients with ditransitive datives in non-native compared to native varieties but also regarding the variability of probabilistic constraints that previous studies report. As pointed out, the majority of corpus-based analyses of the English dative alternation have focused on the ditransitive verb *give* – be it out of convenience for data sampling or due to some other reason. What is more important, however, is the fact that the reported effect of recipient animacy, said to significantly vary in no less than three studies summarised here, was always found in data restricted to the verb *give* (i.e. Bresnan & Hay 2008; Tagliamonte 2014; Szmrecsanyi et al. 2017). Hence, in order to evaluate the regional malleability of

recipient animacy in the present data, I followed the analysis conducted by Bresnan & Hay (2008) using the same model formula in a mixed-effects model but restricting my attention to tokens involving only *give*. No other measures were taken to restrict the dataset further. This GIVE-model includes a random intercept for speaker and fixed effects for recipient givenness, recipient and theme length, recipient and theme pronominality, an interaction of theme givenness with verb semantics and an interaction of recipient animacy and variety (following Bresnan & Hay 2008: 251). VARIETY was again coded with sum coding to compare each variety against the global mean. Since the current dataset does not sample American English, no direct comparison to the findings in Bresnan & Hay (2008) can be made. Nevertheless, the mixed-effects model reveals regional variability of the effect size of recipient animacy in Hong Kong, Indian and Jamaican English. Both graphs in Figure 6.2 plot the likelihood of a prepositional dative (in odds) for the nine varieties depending on whether the recipient is animate (solid line) or inanimate (dashed line). The likelihood of a prepositional dative from the GIVE-model is plotted on the left and can be compared to the effect of recipient animacy in the full model on the right. In the GIVE-model, the effect of recipient animacy on dative choice disappears almost completely in HKE, that is, the likelihood of a prepositional dative is the same whether the recipient is animate or inanimate. In IndE, by contrast, an inanimate recipient increases the likelihood of a prepositional dative far more than in any other variety. The same increase in likelihood can be observed in JamE, where the effect is not as strong as in IndE but still statistically significantly different from the effect of animate recipients on a global level. In comparison, the results of the mixed-effects model on the complete dataset with all verbs reveal no cross-lectal difference in the effect of recipient animacy on the likelihood of a prepositional dative, apart from the fact that inanimate recipients generally increase the probability of a prepositional dative (see Figure 6.2).

While the findings from Bresnan & Hay (2008) cannot be completely confirmed simply on the grounds of the present study's lack of data from American English, the same reasoning does not hold fast for the other two studies that found a statistically significant difference in effect size of recipient animacy across varieties. The difference in the effect of recipient animacy between Canadian and British English observed by Tagliamonte (2014) and Szmrecsanyi et al. (2017) did not show up in the current study. Similarly, the statistically significant differences regarding recipient pronominality between CanE and BrE, regarding verb semantics between BrE and NZE and regarding length between CanE and BrE and between CanE and NZE (Szmrecsanyi et al. 2017)

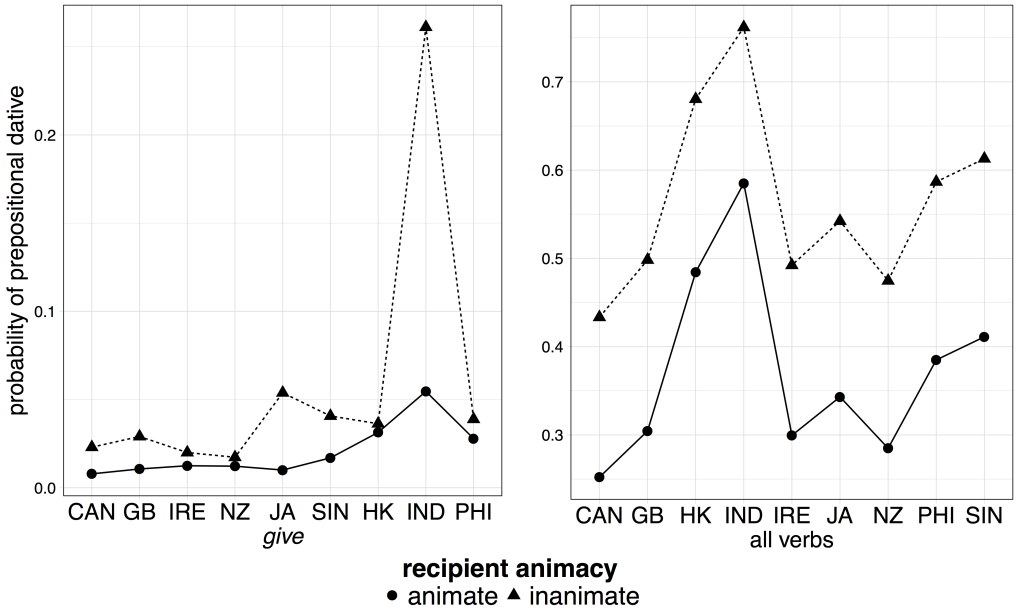


Figure 6.2 (a) left: Variability of effects of recipient animacy across varieties for the verb *give*. **(b) right:** Variability of effects of recipient animacy across varieties for all verbs — Native varieties appear on the left side, non-native varieties appear on the right side of the graphs.

were not observed in the current study. Even when restricting the dataset to *give* and to the most basilectal data possible (spoken informal), the present study did not confirm all findings of previous work. It is thus quite possible that the data analysed is still fairly different from the vernacular speech sampled in Szmrecsanyi et al. (2017) and Tagliamonte (2014) and that coding practices or other restrictions applied to the data confound the picture. (Note, for instance, that Tagliamonte (2014) does not include pronouns in her analysis.)

Apart from the cross-lectal variability of probabilistic constraints, the present study confirms the results of previous work that found length and recipient pronominality to be the most important predictors (e.g. Bresnan & Hay 2008; Schilk et al. 2013). Also, the results of the three separate studies that each focused on one aspect of the regionally malleable probabilistic constraints validate findings of earlier work. As such, structural complexity was shown to be less important than length of the constituents in dative choice (thus confirming the hypothesis by Berlage 2014), lexical effects in the ditransitive dative regarding pronominal recipients find support in language acquisition research (Dodson & Tomasello 1998; McDonough 2006), and the context-dependency of syntactic alternations across different styles also substantiates previous

work (Grafmiller 2014). Besides partly reproducing the results of earlier studies, the current study also offers several innovative advantages over earlier work.

6.4 Innovative aspects

Most basically, the present study adds to current research in linguistics on the descriptive, the methodological and the theoretical level.

On the descriptive level, the present study has patched the gap in our current understanding on the lectal plasticity of probabilistic constraints. By offering an aggregate global perspective that sampled multiple registers and verbs as well as providing detailed micro-analyses of so-called conflict sites (Tagliamonte 2012: 164), my study has confirmed the stability in probabilistic grammars observed in previous work. At the same time, I have also shown that probabilistic variation is ubiquitous in speakers' grammar – be it with regard to the significance or non-significance of particular constraints, differences in effect size or in the relative importance of predictors.

One main surprise of the present analysis is the regional variability of length (see also other studies who found similar effects). So far, length effects or end-weight have been assumed to be typologically robust constraints in that they are connected to general cognitive capacities of the human mind. On the other hand, such variability of length effects might not be that surprising if we consider it as the possible side effect of cognitive indigenisation. If Plan Reuse and Easy First universally apply to language production and comprehension, and if certain (lexically instantiated) constructions or words in a variety are entrenched differently than in other speech communities and varieties, differences in language-internal constraints might be observed. This does not necessarily imply that speakers of, for instance, IndE are in fact more sensitive or less sensitive to length effects. Rather, because such conditioning factors are inherently linked to the linguistic input, differences in linguistic input can lead to differences in the effect strength of conditioning factors. Because length effects and recipient pronominality are the two most important cues in the choice of dative variant, they are also the constraints in which lectal variation can be observed.

On the methodological plane, I have shown how supplementing one's toolbox with multiple statistical techniques enhances our understanding of linguistic phenomena. Thus, by combining mixed-effects models, random forests, collostructional analyses as well as multidimensional scaling, linguistic variation can be explored in all its

facets. While logistic regression gives insights into the effect size of constraints and their possible regional malleability, random forests constitute a solid technique to determine the constraint hierarchy of predictors influencing the choice of variant. Collostructional analyses add to that by offering information on lexical effects, thus providing insights into how lexical and probabilistic factors are interrelated. Multidimensional scaling analysis further offers visualisation and exploratory techniques that highlight the multidisciplinarity of the field of linguistics by providing tools to map varieties in (geographical or even socio-historical) space. MDS compares varieties quantitatively along hitherto unchartered probabilistic dimensions providing objective methodologies for research in comparative sociolinguistics.

In addition to these technical concerns, previous methodological shortcomings of earlier work, that is, most researchers' restrictions on the verb *give* and on a specific register and/or one to two varieties or a specific region, are addressed by the present large-scale comparative study which has sampled data from 86 alternating verbs and 14 diverse genres to investigate patterns of variation in the dative alternation across a total of nine varieties of English.

The supplementary analysis on *give* furthermore highlights that pooling over a large number of verbs – while seemingly concealing potential meaningful differences between varieties – allows for a more comprehensive understanding of the lectal plasticity of the probabilistic grammar underlying the dative alternation. What is more, a large-scale comparative perspective still affords a more fine-grained focus if required. Restricting the focus beforehand on one lexical item – albeit a frequent one – precludes a more aggregate perspective from the start and makes generalisations beyond the particular verb sampled impossible.

Besides the limited focus of previous work on the *give*-alternation, the present study has highlighted the need to complement the spoken vernacular traditionally of interest for variationist sociolinguists with written data and computer-mediated language as well (see also Szmrecsanyi 2017b). Register differences in IndE only became apparent with the addition of the GloWbE data. And it is only by adding more registers to one's dataset that the full extent of the variability of the probabilistic constraints underlying dative choice can be grasped. The analysis has thus shown that REGISTER offers an important aspect to the choice of dative variant (it is a significant factor in HKE, IndE, IrE and JamE) and hence provides an important contextual setting which allows for the variability of stochastic constraints, such as recipient pronominality, across different registers.

On the theoretical plane, the research adopted here has examined linguistic variation from multiple angles, modelling over large-scale datasets as well as conducting closer syntactic and semantic analyses of particular lexical items and constructions. The present study thus fits in tightly with a variation-centred, usage- and experienced-based probabilistic grammar approach as well as with recent research in Cognitive Sociolinguistics. The presented study is furthermore also consonant with principles of frequency-based and exemplar-based approaches to language variation. Common to all of these approaches is the assumption that grammar is the "cognitive organization of one's experience with language" (Bybee 2006: 711) and that probabilistic knowledge is derived from language experience. More specifically, my findings tie in with recent research in Cognitive Sociolinguistics which view variation in language from a cognitive as well as socio-cultural perspective. Cognitive Sociolinguistic approaches combine the commitment of Cognitive Linguistics with the social dimension of language variation inherent to sociolinguistic accounts. In that regard, the results of my study highlight that speakers' grammatical knowledge can only be understood when socially contextualised in language usage. In turn, these contextualisations (for instance, within a regionally defined speech community) result in the acquisition of social and not just typical cognitive constraints on language variation, indicating that a strict separation between cognitive and social constraints might not necessarily be possible. Similarly to cognitive constraints, social constraints on linguistic variation do not have a categorical impact on language but are stochastically derived from a speaker's experience with language and the context language is used in (formality of situation, social characteristics of speaker, and so on) (Foulkes & Docherty 2006; Geeraerts 2010a; see also Rosseel 2017 for an in-depth study on the social meaning of variation). Cognitive indigenisation thus does not only refer to the indigenisation of cognitive, language-internal constraints but to the simultaneous integration of contextual (social) constraints in language variation as well. A cognitive sociolinguistic approach can account for cognitive indigenisation being not only regionally (or speech community)-dependent but overall context-dependent – that is, across different registers, lexical preferences or social groups – since the contextual parameters provide the background for the indigenisation process.

Context-dependency across social groups might seem a bit far fetched considering that the dative alternation is a syntactic alternation and as such has not been traditionally considered a sociolinguistic variable – that is, a variable that expresses social meaning (see Lavandera 1978). However, the context-dependency of linguistic varia-

tion also holds here: Recent research has shown, for instance, that there seems to be a gender effect in the use of dative variant. Theijssen et al. (2011) find Australian males to prefer the prepositional dative more than Australian female speakers. Jenset et al. (2017) report the same effect for British English. Similarly, my own data indicates an increased use of prepositional dative variants among male speakers compared to female speakers, especially in Irish and Jamaican English. This preference of female speakers for the ditransitive dative might very well be a reflection of females leading the change towards more ditransitive – a trend that has been observed in apparent time by Tagliamonte (2014) in both British and Canadian English (sampling over multiple verbs) and by Grimm & Bresnan (2009) in British and American English journalistic prose. This female preference for ditransitive, however, also contrasts partly with findings in Tagliamonte (2014) who shows females to prefer the prepositional dative across all age groups in Canadian English. Also controversially, Bresnan & Hay (2008) report a preference of younger and older speakers for the prepositional dative while middle-aged speakers use the ditransitive dative more often (note that Bresnan & Hay 2008 focus on *give*).

Together, these findings emphasise that the emergence of cognitive indigenisation can only be adequately recognised if we take both the social as well as cognitive nature of language into account.

6.5 Challenges

Finally, a caveat is in order here. To test the cognitive plausibility of statistical models such as the current one, corpus-based analyses have been comparing the models' performance with the prediction accuracy of native speakers obtained in experimental settings (see Klavan & Divjak 2016: 357). Even though such studies show that language users' implicit knowledge of variation patterns reflects on the whole the usage probabilities attained from statistical models much more closely than expected (e.g. Bresnan et al. 2007a; Bresnan & Ford 2010), this is not always the case. Comparisons often reveal marginal but existing differences between observational aggregate data and behavioural individual data. We thus have to be circumspect when drawing conclusions about speakers' linguistic knowledge based on the results from regression models. While regression techniques might not necessarily mirror the cognitive reality in speakers' mind with 100% accuracy, they can still be used to assess the relative weighting of simultaneously interacting constraints on language

performance and are thus valid and cognitively realistic approximations (Klavan & Divjak 2016: 379). Truly cognitive models (for instance, memory-based learning, naïve discriminative learning) are currently being developed (see Milin et al. 2016) and they will certainly enhance our grasp of speakers' grammatical knowledge in future work.

Another challenge that this study encountered is related to the data sources tapped into. Comparisons carried out with earlier work that focused on more basilectal speech are problematic in view of the acrolectal data sampled in ICE and GloWbE. Also, the limited information available on speakers' background makes it harder to incorporate social factors apart from region into the analysis. Missing information on speakers is especially problematic in GloWbE (see Davies & Fuchs 2015) where texts were scrambled off the internet using the site's URL as indication of regional origin. This can of course not prevent people of, say Indian origin, posting on blogs in the UK, which in turn confounds possible regional differences between varieties. The possible lack of regional variability in GloWbE can be empirically verified to some extent by the proportional distributions of ditransitive and prepositional datives across the nine national varieties. In ICE, the data from IndE contains significantly more prepositional datives than all other varieties (with the exceptions of maybe HKE and PhiE) while the distribution is more equal in GloWbE in that respect (see Figure 6.3). Apart from the difference in IndE, however, no other statistically significant differences between GloWbE and the aggregate of ICE were observed in the mixed-effects model. Results from 'small and tidy' corpora such as ICE thus largely match findings from 'big and messy' corpora such as GloWbE (see Hundt & Leech 2012). It is only when we take a closer look at the individual registers that differences emerge.

Finally, on a more general level, the aggregate perspective adopted here can overlook more fine-grained differences in the pattern of variation related to specific verbs or recipients, to registers or social groups or other contextual aspects that I have not touched upon. What is more, an aggregate perspective makes generalisations on the community level possible but says little to nothing about individual-level variation. While individual-level variation is only marginally relevant for (Labovian) Variationist Sociolinguistics and comparatively important for Cognitive Sociolinguists (Walker & Meyerhoff 2013: 175; see also Geeraerts 2010b), the possible discrepancy between community-level patterns and individual-level patterns remain an issue that the present analysis has not addressed so far. Cognitive indigenisation has mainly been defined in terms of community-level processes, assuming that what is happening

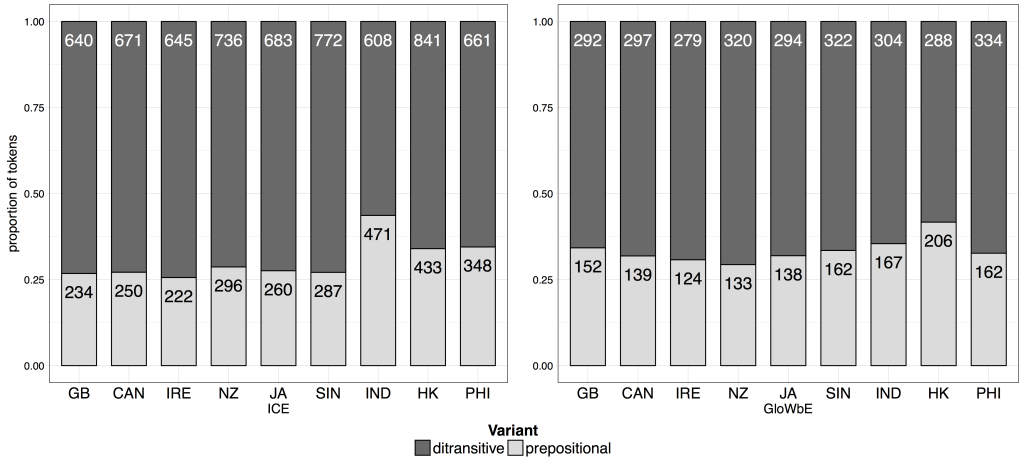


Figure 6.3 Proportional distribution of ditransitive and prepositional datives in ICE (right) and GloWbE (left) — Native varieties appear on the left side, non-native varieties appear on the right side of the graphs.

in the individual can be aggregated to the larger whole. However, the diversity of data sampled in ICE – to come back to the methodological issues – calls into question the extent to which the community-level patterns observed are homogeneously distributed among the individual speakers. That such a homogeneous speech community might not always exist in the context of ICE and GloWbE is shown in Figure 6.4. Figure 6.4 plots the random intercepts by speaker from by-variety mixed-effects models following the model formula in (73) in Section 5.7. The models include the five most important predictors given the conditional random forest and a random intercept for lexical effects and speaker. Ireland is not plotted since the model with a random intercept by speaker did not converge.

As the visualisation of random intercepts illustrate (see Figure 6.4), the data from BrE and SinE is much more homogeneous and coherent than the data from PhiE, HKE, IndE and CanE. In SinE, the random intercept of speaker even accounts for zero variance. This difference in intradialectal homogeneity between varieties could potentially be ascribed to register effects, something which certainly warrants further exploration.

Hence, while it is true that a micro-perspective might obscure the more generalisable patterns, it is also true that an aggregate perspective can conceal the even more subtle patterns of variability to be found in the data. Even though an aggregate perspective offers new insights into the cognitive underpinnings of the English dative

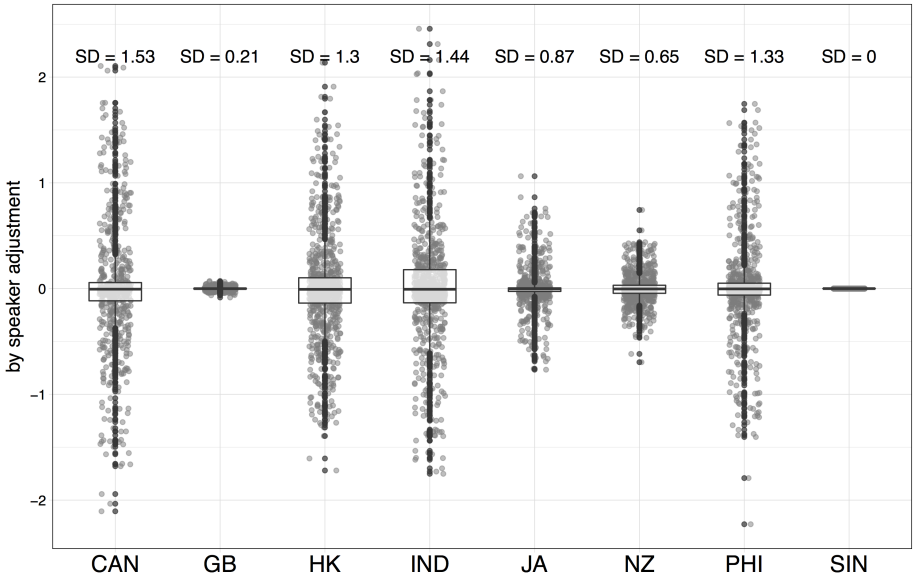


Figure 6.4 By-speaker adjustments to the intercept of a model with the five most important predictors — Inter-individual variation is larger in CanE, HKE, IndE, PhiE compared to the other varieties. Varieties are ordered alphabetically.

alternation on the whole, there still remains much to be explored on the micro-level of variation in syntactic alternations in English more generally and in the dative alternation more specifically.

7

Conclusion

This thesis has investigated regional variation in probabilistic grammars focusing on the alternation between the ditransitive (e.g. *John gives Mary the apple*) and the prepositional dative (e.g. *John gives the apple to Mary*) across nine national varieties of English. Sampling dative observations of naturalistic language use from spoken and written registers (ICE and GloWbE) and including a total of 86 alternating dative verbs, this study is the first to offer a comprehensive overview of regional variation in the probabilistic constraints and the lexical effects that drive the English dative alternation. The results of various multifactorial statistical analyses (mixed-effects, random forests, collexeme analyses) show that the predictors recipient pronominality and length have a diverging impact on dative choice in Hong Kong, Indian, Irish and Jamaican English compared to the global average. Recipient pronominality and length are thus the two factors most malleable to probabilistic indigenisation (a term that designates the gradual shifts in the stochastic constraints that drive linguistic variation) and they are also the two most important factors across all nine varieties as evidenced by the random forest. Further probing into these two factors highlighted first, that the length of the constituents in characters is a better predictor of dative choice than the constituents' syntactic complexity. What is more, length is regionally malleable while syntactic complexity is not, even when the dataset is restricted to nominal constituents only. Second, the extent to which pronominal recipients are associated with the ditransitive and the prepositional dative differs across varieties: Speakers of non-native varieties are more likely to use the pronominal recipient in the ditransitive dative than native speakers of English. The mixed-effects model

also showed a significant effect of CORPUS in Indian English, which was taken as a starting point to explore the register-specificity of the dative alternation further. The analyses in this third step showed that register impacts dative choice significantly in Hong Kong, Indian, Irish and Jamaican English. What is more, the prepositional dative is overall more likely in all five registers in non-native varieties compared to native varieties. A closer look at intra-systemic variation within each register finally revealed that first and foremost recipient pronominality differs in its effect size across different registers, that is, recipient pronominality is not only regionally but also stylistically variable and thus truly a cross-lectally malleable constraint. In a final step, methods developed in Comparative Sociolinguistics were employed to quantify the probabilistic distance between varieties' grammar and thus to assess the limits of cross-varietal variation. Probabilistic distances were compared along three different dimensions (statistical significance, effect size, constraint hierarchy). In this final step, the measured probabilistic distance between varieties led to an overall stability score for each dimension. Subsequent visualisation with MDS offers no conclusive distinction between varieties based on variety type (native vs. non-native) or socio-historical grounds (Asian Varieties, or varieties clustered based on their evolutionary development). That being said, the analysis does set apart North American (influenced) varieties (CanE, PhiE) from the rest. The results suggest that the three dimensions gauge different – maybe even unrelated – aspects of speakers' probabilistic grammar. All in all, the results highlight that probabilistic grammars are not as stable as was hitherto believed to be the case (see Bernaisch et al. 2014) and that their presumed stability or variability is dependent on the lexical items and the syntactic alternation included in one's analysis.

By digging deep enough into speakers' grammatical knowledge, the present study has shown that alleged subtle differences in probabilistic constraints can have their cause in structural differences in lexical preferences between varieties. The presumption of a stable probabilistic grammar has also been refuted in the exploration of variability in probabilistic grammars using comparative sociolinguistic methods. Fundamentally, the study draws on the concept of probabilistic indigenisation to account for the gradient shifts in the probabilistic grammar of regionally distinct speech communities. Since the concept of indigenisation is especially prolific in the field of World Englishes, the present study adds to scholarship in both Cognitive Sociolinguistics and World Englishes by bringing these two research paradigms closer together. With regard to World Englishes, my study has shown that processes of

indigenisation or nativisation do not only take place at the interface between the lexicon and morphosyntax but also in the probabilistic constraints that fuel variation in language. By including a cognitive dimension in the analysis of variation across varieties of English, we can situate the observed variation within the broader context of language comprehension and production. Probabilistic indigenisation on the other hand eventually leads to cognitive indigenisation, that is, the creation of different lects that are distinct from each other on the underlying probabilistic level. Regarding research in Cognitive Sociolinguistics then, the current work has highlighted the ubiquity of linguistic variation in speakers' probabilistic grammar(s) and has elaborated on the causes of such lectal variation. What is more, the results of the present study have raised questions about the nature and scope of stability in probabilistic grammars on a community as well as individual level. Questioning the stability of individual's probabilistic grammar(s) carries implications for both Cognitive Sociolinguistics and Variationist Sociolinguistics as the presented results challenge the well-formed boundaries of lects. If probabilistic lectal variation is indeed ubiquitous and if lects thus cannot be clearly defined on a probabilistic level, aggregating over individuals' probabilistic grammar to a community grammar becomes contentious. Such an aggregation might be less problematic in the morphosyntactic and lexical domain but raises issues for Cognitive Sociolinguistics (and to some extent Variationist Sociolinguists) who are particularly interested in the relationship between individuals' grammar and community-level grammars and who centre their attention on the stochastic constraints that drive variation. If an individual's grammar does not constitute a coherent unit, aggregating over individuals' grammar will obscure more linguistic variation than hitherto assumed.

The study was faced with two challenges which limits the comprehensiveness of the analysis to some extent. First, the current study did not consider any social parameters in its analysis of speakers' choice between the two dative variants. Syntactic alternations such as the dative alternation, while claimed to carry no differences in propositional meaning, are not known to carry social meaning either. This is, for instance, not the case with phonological variation (e.g. [ɪn]) vs. [ɪŋ]) where the two phonological variants are undisputedly semantic equivalent and, also undisputedly, socially variable. The extent to which the traditional analysis of phonological variants' social meaning can be extended to other linguistic domains, for instance, the lexicon or syntax, has been repeatedly debated (see Lavandera 1978). Most studies on syntactic alternations thus refrain from including social parameters in their analysis (such as

speaker sex or age) since some researchers claim that the two syntactic variants are already so different in propositional meaning that these meaning differences oust any social meaning attached to the variants. As the short interlude into gender differences in the Discussion has shown, however, dative variants are differently distributed by gender. Whether these distributional differences are reflected in differences in the probabilistic domain remains to be investigated.

Second, the lack of speaker information in GloWbE and the small number of tokens by individual speakers in ICE (and presumably GloWbE), left the question unresolved regarding the relationship between individual and community grammars. While it is clearly desirable of future work to include demographic factors in their analysis – as illustrated above – we should also start sampling large quantities of data from individual speakers in order to connect individual-level variation with community-level variation. Putting this methodological issue on our research agendas will enable us to not only find additional supporting evidence for social meaning attached to syntactic variants. By focusing on the interplay between individual- and community-level variation, future research might also be able to pinpoint the lower floor of lectal variation within the individual and to explore intra-individual lectal variation in more detail.

Besides these two challenges, future work could also increase our knowledge of regional variation in syntactic alternations by extending the analysis to a more comprehensive and focused empirical investigation of onomasiological variation and semasiological variation in the dative alternation. With respect to onomasiological variation, future work could include additional variants, for instance the beneficiary construction or dialectal variants that have been shown to alternate with either of the two standard dative variants as well. Such a comprehensive analysis would follow the *Principle of Accountability* more closely by including all possible "alternate ways of saying 'the same' thing" (Labov 1972a: 188). Regarding semasiological variation, the current study has already offered some insights into regional semasiological variation by comparing the variants' lexical profiles across varieties. The semantic coherence between the two variants is clearly something that deserves a more empirical investigation, for instance by using random intercepts from by-variety models to find lexical distance between varieties, by investigating variability in the effect of language-internal constraints by verb sense and hence the meaning associated with a variant or by adding semantic vector space models to the researcher's toolkit. Semantic vector space models could assess meaning differences between variants based on

the verbs, recipients and themes (as well as the subject of the clause) used in each variant. Regional variation can then be explored by comparing clusters of variants that instantiate the same meaning across varieties. Extending the semasiological perspective of the present study with semantic vector space models would address questions of how many meanings are associated with the ditransitive or prepositional dative and provide insights into meaning differences across varieties or registers.

One other desirable asset of future work is to test the cognitive reality of the probabilistic constraints measured in statistical models (see Klavan & Divjak 2016). Conducting rating task experiments in the spirit of Bresnan & Ford (2010) would enable us to not only compare the models' performance with the prediction accuracy of speakers but also to test the cognitive plausibility of statistical models. Work is under way to conduct such experiments right now, offering more insights into speakers' knowledge of probabilistic grammars in the near future.

Lastly, it remains desirable of future research to take an exhaustive look at substrate languages and their influence on grammatical variation that goes beyond the discussion presented in Chapter 6. The impact of substrate languages on the choice of linguistic variants could, for instance, be assessed systematically based on the information provided in the *World Atlas of Language Structure* (WALS) (Dryer & Haspelmath 2013) (as done in Heller 2018) or by drawing on existing research on the use of datives in languages that constitute substrates to national varieties of English (for instance, Hindi or Cantonese).

Despite all limitations and challenges for future work, and even though much remains to be investigated in the English dative alternation, the present study hopefully represents the first step in a line of research that takes a more comprehensive stance to the analysis of grammatical variation in syntactic alternations. The amount and diversity of data used – the wide range of verbs, registers and varieties – is unprecedented in earlier work. Only by adopting such an aggregate perspective, does a comprehensive and detailed investigation of regional variation in probabilistic grammars become possible.

Appendix A

Table A The 16 levels of structural complexity following Berlage (2014) (to be continued on the next page)

Code	Category	Comments	Examples
's'	simple	any pronoun or NP with [(Det) (A) N] structure	*subscriptions, any old rubbish, her head, anyone else, the name Bender*
'co'	coordinated NP	noun phrases involving multiple heads joined with *and, or, but, though, along with*, etc.	*the onions and the potatoes, Accounting or Economics, silt and floodwaters*
'pp'	prepositional phrase	any PP that is unambiguously modifying the constituent NP (and in cases of subordination, also the verb phrase of the subordinate clause); this includes all *of*-PPs	*the lies about Obama, research on these writers, that line of work, a picture with a frame*
'postad'	Adj/Adv/Det	NP with postmodifying adjective/adverb/determiner in [NP Adj/Det/Adv] structure	*the people there, a day off, the juiciest steak ever*

Code	Category	Comments	Examples
'gn'	genitive	NP with an *s-* genitive	*my father's gun*
'appnom'	nominal appositions	NPs that postmodify the head by specifying/restricting it	*Jill, Yeat's daughter; Angela, his wife*
'ge'	general extenders	general extenders following the head and which are not coordinating two heads	*a story and something*
'nonfin'	nonfinite clauses	NP with nonfinite clauses, such as *-ing/-ed/to*, also called 'vp' in the shared annotation	*a chance to complain, a letter containing.., people injured on the streets*
'rc'	(finite) relative clause	finite, restrictive relative clauses. These can have overt or null relative pronouns.	*the guy that caused the accident, the toys you thought were our favorites*
'cp'	complement clause	NP with clausal complement, i.e. overt or null *that*-clauses	*the fact that I had seen him, the impression gambling was acceptable*
'nc'	nominal clauses	clauses that substitute a noun; they often occur on the 2nd level of postmodification in a [pp nc] structure or follow a verb	*a lesson on what constitutes an offence, a picture of how it was used, saying that he should use it*
'advc'	adverbial clauses	clauses that substitute an adverb	*before I left, when talking to yourself*

Appendix B

Two analyses were performed in order to assess concept validity and concept reliability of the three methods proposed by Comparative Sociolinguistics.

Concept validity

Heller's (2018) simulation study uses four predictors and their typical coefficient estimates derived from a regression model to generate new data for 10 fictitious varieties. For each of these varieties, the degree of grammatical variability was increased in 11 stages by always 10% starting from 0% to 100% around the original coefficient estimate. So, each variety has a stage where the coefficient estimates are equivalent to the original coefficient estimates. For instance, the coefficient estimate for possessor animacy is 4 in the original model which is equivalent to the 0% condition, i.e. all varieties have a coefficient estimate of 4 in the 0% condition. In the next stage, the 10% condition, coefficient estimates are allowed to have a standard deviation of up to 10% of their original value. Coefficient estimates were thus randomly generated. Variability in coefficient estimates is then subsequently increased for all 10 varieties until the 100% condition is reached. Once the coefficient estimates have been computed, 500 observations are generated for each of the 10 varieties and for each of the variability conditions ($N = 55{,}000$ observations) by calculating the predicted outcome based on the computed coefficient estimates in each variety and condition. Next, regression models and random forests are fitted to the dataset separately by each variety and condition and the output of the models is compared along the three lines of evidence: statistical significance, effect size and constraint ranking. For each line, a similarity score is calculated which assesses the similarity between the 10 varieties in one of the 11 conditions. The results of that simulation study show that all three lines pick up on increasing variability in the data: The higher the variability (i.e. the more the condition moves from 0% to 100%), the lower the

similarity measured between the nine varieties.

Concept reliability

The results of the bootstrapping study, which was fitted to assess concept reliability of the three lines of evidence, are shown in Figure 7. Figure 7 illustrates that the coefficient estimates (i.e. effect sizes) from the mixed-effects models provide the most reliable measure with very small confidence intervals. However, they also provide the lowest stability score in comparison.

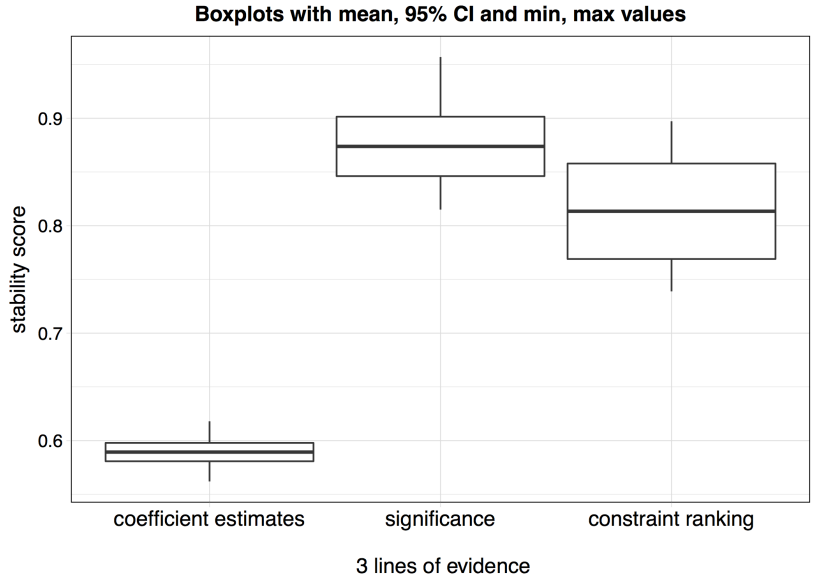


Figure A Bootstrapping to assess concept reliability of comparative sociolinguistic methods — The comparison across the boxplots highlights that coefficient estimates provide the most reliable measure but also the one with the lowest stability score.

References

Aarts, Bas, Sean Wallis & Sidney Greenbaum. 1998. ICE-GB: The British component of the International Corpus of English.

Aissen, Judith. 2003. Differential object marking: Iconicity vs. economy. *Natural Language and Linguistic Theory* 21(3). 435–483.

Aldenderfer, Mark S. & Roger K. Blashfield. 1984. *Cluster analysis*. Newbury Park/London/New Delhi: SAGE Publications.

Allen, Cynthia L. 1995. *Case marking and reanalysis*. Oxford: Oxford University Press.

Allen, Cynthia L. 2006. Case syncretism and word order change. In Ans Van Kemenade & Bettelou Los (eds.), *The handbook of the history of English*, 201–223. Oxford: Blackwell.

Aloni, Maria & Floris Roelofsen. 2012. Interpreting concealed questions. *Linguistics and Philosophy* 34(5). 443–478.

Anderwald, Lieselotte. 2004. The varieties of English spoken in the southeast of England: Morphology and syntax. In Bernd Kortmann, Edgar W. Schneider, Kate Burridge, Rajend Mesthrie & Clive Upton (eds.), *A handbook of varieties of English*, Vol. 2: Morphology and syntax, 175–195. Berlin/New York: Mouton de Gruyter.

Arnold, Jennifer, Anthony Losongco, Thomas Wasow & Ryan Ginstrom. 2000. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language* 76(1). 28–55.

Aston, Guy & Lou Burnard. 1998. *The BNC handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

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

Bailey, Robert. 2013. The quantitative paradigm. In J. K. Chambers & Natalie Schilling-Estes (eds.), *The handbook of language variation and change*, 85–107. Oxford: Wiley-Blackwell.

Baker, Carl Lee. 1979. Syntactic theory and the projection problem. *Linguistic Inquiry* 10(4). 533–581.

Bartoń, Kamil. 2016. *MuMIn: Multi-Model Inference*. Available at: https://CRAN.R-project.org/package=MuMIn.

Barðdal, Jóhanna, Kristian Emil Kristoffersen & Andreas Sveen. 2011. West Scandinavian ditransitives as a family of constructions: With a special attention to the Norwegian 'V - REFL - NP' construction. *Linguistics* 49(1). 53–104.

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

Bauer, Laurie & Winifred Bauer. 2002. Can we watch regional dialects developing in colonial English? The case of New Zealand. *English World-Wide* 23(2). 169–193.

Bauer, Laurie & Paul Warren. 2008. New Zealand English: Phonology. In Kate Burridge & Bernd Kortmann (eds.), *Varieties of English*, Vol. 3: The Pacific and Australasia, 39–63. Berlin/New York: Mouton de Gruyter.

Beckford Wassink, Alicia. 1999. Historic low prestige and seeds of change: Attitudes toward Jamaican Creole. *Language in Society* 28(1). 57–92.

Behaghel, Otto. 1909. Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen* 25. 110–142.

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

Berlage, Eva. 2014. *Noun phrase complexity in English*. Cambridge: Cambridge University Press.

Bernaisch, Tobias, Stefan Th. Gries & Joybrato Mukherjee. 2014. The dative alternation in South Asian English(es): Modelling predictors and predicting prototypes. *English World-Wide* 35(1). 7–31.

Bhatt, Rakesh M. 2004. Indian English: Syntax. In Bernd Kortmann, Edgar W. Schneider, Kate Burridge, Rajend Mesthrie & Clive Upton (eds.), *A handbook of varieties of English*, Vol. 2: Morphology and syntax, 1016–1030. Berlin/New York: Mouton de Gruyter.

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

Boberg, Charles. 2008. English in Canada: Phonology. In Edgar Schneider (ed.), *Varieties of English*, Vol. 2: The Americas and the Caribbean, 144–160. Berlin/New York: Mouton de Gruyter.

Bock, Kathryn. 1982. Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review* 89(1). 1–47.

Bock, Kathryn. 1986. Syntactic persistence in language production. *Cognitive Psychology* 18(3). 355–387.

Bock, Kathryn & Zenzi M. Griffin. 2000. The persistence of structural priming: Transient activation or implicit learning? *Journal of Experimental Psychology: General* 129(2). 177–192.

Bock, Kathryn & David Irwin. 1980. Syntactic effects of information availability in sentence production. *Journal of Verbal Learning and Verbal Behavior* 19(4). 467–484.

Bod, Rens, Jennifer Hay & Stefanie Jannedy (eds.). 2003. *Probabilistic linguistics*. Cambridge, MA: MIT Press.

Bolt, Philip & Kingsley Bolton. 1996. The International Corpus of English in Hong Kong. In Sidney Greenbaum (ed.), *Comparing English worldwide: The International Corpus of English*, 197–214. Oxford/New York: Clarendon Press/Oxford University Press.

Bolton, Kingsley & Joseph Hung. 2006. The ICE Hong Kong Corpus. Available at: http://ice-corpora.net/ice/download.htm.

Branigan, Holly P., Martin J. Pickering & Alexandra Cleland. 2000. Syntactic co-ordination in dialogue. *Cognition* 75(2). B13–25.

Bresnan, Joan. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Sam Featherston & Wolfgang Sternefeld (eds.), *Roots: Linguistics in search of its evidential base*, 75–96. Berlin/New York: Mouton de Gruyter.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina & R. Harald Baayen. 2007a. Predicting the dative alternation. In Gerlof Boume, Irene Kraemer & Joost Zwarts (eds.), *Cognitive foundations of interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bresnan, Joan, Ashwini Deo & Devyani Sharma. 2007b. Typology in variation: A probabilistic approach to *be* and *n't* in the Survey of English Dialects. *English Language and Linguistics* 11(2). 301–346.

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

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

Bresnan, Joan & Tatiana Nikitina. 2003. On the gradience of the dative alternation. Manuscript.

Bresnan, Joan & Tatiana Nikitina. 2009. The gradience of the dative alternation. In Linda Uyechi & Lian Hee Wee (eds.), *Reality exploration and discovery: Pattern interaction in language and life*, 161–184. Stanford: CSLI Publications.

Bruyn, Adrienne, Pieter Muysken & Maaike Verrips. 1999. Double-object constructions in the creole languages: Development and acquisition. In Michel DeGraff (ed.), *Language creation and language change: Creolization, diachrony and development*, 329–373. Cambridge, MA: MIT Press.

Buschfeld, Sarah, Thomas Hoffmann, Magnus Huber & Alexander Kautzsch (eds.). 2014. *The evolution of Englishes: The Dynamic Model and beyond*. Amsterdam/Philadelphia: John Benjamins.

Buschfeld, Sarah & Alexander Kautzsch. 2017. Towards an integrated approach to postcolonial and non-postcolonial Englishes. *World Englishes* 36(1). 104–126.

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

Bybee, Joan & Paul Hopper. 2001. *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins.

Bürkle, Daniel. 2015. *The acquisition of sentence alternations: How children understand and use the English dative alternation*. Christchurch: University of Canterbury PhD dissertation.

Campbell, Aimee L. & Michael Tomasello. 2001. The acquisition of English dative constructions. *Applied Psycholinguistics* 22(2). 253–267.

Cedergren, Henrietta J. 1973. *The interplay of social and linguistic factors in Panama*. Cornell: Cornell University PhD dissertation.

Cedergren, Henrietta J. & David Sankoff. 1974. Variable rules: Performance as a statistical reflection of competence. *Language* 50(2). 333–355.

Chambers, J. K. 2009. *Sociolinguistic theory: Linguistic variation and its social significance*. Malden/Oxford: Wiley-Blackwell.

Chambers, J. K. 2012. Homogeneity as a sociolinguistic motive in Canadian English. *World Englishes* 31(4). 467–477.

Childers, Jane B. & Michael Tomasello. 2001. The role of pronouns in young children's acquisition of English transitive construction. *Developmental Psychology* 37(6). 739–748.

Christie, Pauline. 2003. *Language in Jamaica*. Kingston: Arawak Publications.

Colleman, Timothy & Bernard De Clerck. 2009. 'Caused motion'? The semantics of the English *to*-dative and the Dutch *aan*-dative. *Cognitive Linguistics* 20(1). 5–42.

Colleman, Timothy & Bernard De Clerck. 2011. Constructional semantics on the move: On semantic specialization in the English double object construction. *Cognitive Linguistics* 22(1). 183–209.

Collins, Peter. 1995. The indirect object construction in English: An informational approach. *Linguistics* 33(1). 35–49.

Conwell, Erin & Katherine Demuth. 2007. Early syntactic productivity: Evidence from dative shift. *Cognition* 103(2). 163–179.

Crawford, Jarret T., Lee Jussim & Jane M. Pilanski. 2014. How (not) to interpret and report main effects and interactions in multiple regression: Why Crawford and Pilanski did not actually replicate Lindner and Nosek (2009). *Political Psychology* 35(6). 857–862.

Cueni, Anna. 2004. Predicting the outcome of the choice between the dative constructions of English. Manuscript.

Davies, Mark. 2013. *Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries*. Available at: http://corpus.byu.edu/glowbe/.

Davies, Mark & Robert Fuchs. 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.

De Cuypere, Ludovic. 2010. A discourse-based account of the Old English double object alternation. *Sprachwissenschaft* 35(3). 337–368.

De Cuypere, Ludovic. 2015a. A multivariate analysis of the Old English ACC + DAT double object alternation. *Corpus Linguistics and Linguistic Theory* 11(2). 225–254.

De Cuypere, Ludovic. 2015b. The Old English *to*-dative. *English Language and Linguistics* 19(1). 1–26.

De Cuypere, Ludovic, Evelyn De Coster & Kristof Baten. 2014. The acquisition of the English dative alternation by Russian foreign language learners. *Phrasis (Gent): Studies in language and literature* 2. 187–212.

De Cuypere, Ludovic & Saartje Verbeke. 2013. Dative alternation in Indian English: A corpus-based analysis. *World Englishes* 32(2). 169–184.

de Marneffe, Marie-Catherine, Scott Grimm, Inbal Arnon, Susannah Kirby & Joan Bresnan. 2012. A statistical model of the grammatical choices in child production of dative sentences. *Language and Cognitive Processes* 27(1). 25–61.

Deuber, Dagmar. 2014. *English in the Caribbean: Variation, style and standards in Jamaica and Trinidad*. Cambridge: Cambridge University Press.

Dodson, Kelly & Michael Tomasello. 1998. Acquiring the transitive construction in English: The role of animacy and pronouns. *Journal of Child Language* 25(3). 605–622.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available at: http://wals.info/.

Ehret, Katharina. 2008. *Analyticity and syntheticity in East African English and British English: A register comparison*. Freiburg: Albert-Ludwigs-Universität Freiburg i. Br. BA thesis.

Elenbaas, Marion B. 2013. The synchronic and diachronic status of English light verbs. *Linguistic Variation* 13(1). 48–80.

Ellis, Nick C. 2002. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition* 24(2). 143–188.

Ellis, Nick C. 2006. Selective attention and transfer phenomena in L2 acquisition: Contigency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics* 27(2). 164–194.

Erteschik-Shir, Nomi. 1979. Discourse constraints on dative movement. In Talmy Givón (ed.), *Discourse and syntax*, 441–467. New York: Academic Press.

Farquharson, Joseph T. 2013. Jamaican structure dataset. In Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath & Magnus Huber (eds.), *Atlas of Pidgin and Creole Language Structures Online*, Leipzig: Max Planck Institute for Evolutionary Anthropology. Available at: http://apics-online.info/contributions/8.

Fasold, Ralph. 1984. *The sociolinguistics of society*. Oxford: Blackwell.

Feagin, Crawford. 1979. *Variation and change in Alabama English: A sociolinguistic study of the white community*. Washington, DC: Georgetown University Press.

Ferreira, Fernanda. 1991. Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language* 30(2). 210–233.

Ferreira, Fernanda. 1994. Choice of passive voice is affected by verb type and animacy. *Journal of Memory and Language* 33(6). 715–736.

Filppula, Markku. 1999. *The grammar of Irish English: Language in Hibernian style*. London/New York: Routledge.

Foulkes, Paul & Gerard Docherty. 2006. The social life of phonetics and phonology. *Journal of Phonetics* 34(4). 409–438.

Fox, John. 2003. Effect displays in R for generalised linear models. *Journal of Statistical Software* 8(15). 1–27. Available at: http://www.jstatsoft.org/v08/i15/.

Gahl, Susanne & Susan Garnsey. 2004. Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language* 80(4). 748–775.

Garretson, Gregory, M. Catherine O'Connor, Barbora Skarabela & Marjorie Hogan. 2004. Coding practices used in the project "Optimality Typology of Determiner Phrases". Available at: http://npcorpus.edu/documentation/index.html.

Gast, Volker. 2007. *I gave it him*: On the motivation of the 'alternative double object construction' in varieties of British English. *Functions of Language* 14(1). 31–56.

Geeraerts, Dirk. 2005. Lectal variation and empirical data in Cognitive Linguistics. In Francisco J. Ruiz de Mendoza Ibañez & M. Sandra Peña Cervel (eds.), *Cognitive Linguistics: Internal dynamics and interdisciplinary interactions*, 163–189. Berlin/New York: Mouton de Gruyter.

Geeraerts, Dirk. 2010a. Recontextualizing grammar: Underlying trends in thirty years of Cognitive Linguistics. In Elzbieta Tabakowska, Michal Choinski & Lukasz Wiraszka (eds.), *Cognitive Linguistics in action: From theory to application and back*, 71–102. Berlin/New York: Mouton de Gruyter.

Geeraerts, Dirk. 2010b. Schmidt redux: How systematic is the linguistic system if variation is rampant? In Kasper Boye & Elisabeth Engberg-Pedersen (eds.), *Language usage and language structure*, 237–262. Berlin/New York: Mouton de Gruyter.

Geeraerts, Dirk, Stefan Grondelaers & Peter Bakema. 1994. *The structure of lexical variation: A descriptive framework for cognitive lexicology*. Berlin: Mouton de Gruyter.

Geeraerts, Dirk, Gitte Kristiansen & Yves Peirsman. 2010. *Advances in Cognitive Sociolinguistics*. Berlin/New York: Mouton de Gruyter.

Gelman, Andrew. 2008. Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine* 27(15). 2865–2873.

Gelman, Andrew & Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Gerwin, Johanna. 2013. *Give it me*!: Pronominal ditransitives in English dialects. *English Language and Linguistics* 17(3). 445–463.

Gerwin, Johanna. 2014. *Ditransitives in British English dialects*. Berlin/New York: Mouton de Gruyter.

Goldberg, Adele E. 1995. *Constructions: A Construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Goldberg, Adele E. 2003. Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences* 7(5). 219–224.

Gordon, Elizabeth & Margaret Maclagan. 2008. Regional and social differences in New Zealand: Phonology. In Kate Burridge & Bernd Kortmann (eds.), *Varieties of English*, Vol. 3: The Pacific and Australasia, 64–76. Berlin/New York: Mouton de Gruyter.

Grafmiller, Jason. 2014. Variation in English genitives across modality and genres. *English Language and Linguistics* 18(3). 471–496.

Grafmiller, Jason, Melanie Röthlisberger, Benedikt Heller & Benedikt Szmrecsanyi. 2016. Annotation of common features for the genitive, dative, and particle placement alternations. Manuscript.

Green, Georgia. 1974. *Semantics and syntactic regularity*. Bloomington: Indiana University Press.

Greenbaum, Sidney. 1988. A proposal for an international computerized corpus of English. *World Englishes* 7(3). 315.

Greenbaum, Sidney (ed.). 1996a. *Comparing English worldwide: The International Corpus of English*. Oxford/New York: Clarendon Press/Oxford University Press.

Greenbaum, Sidney. 1996b. Introducing ICE. In Sidney Greenbaum (ed.), *Comparing English worldwide: The International Corpus of English*, 3–12. Oxford/New York: Clarendon Press/Oxford University Press.

Gries, Stefan Th. 2001. A multifactorial analysis of syntactic variation: Particle movement revisited. *Journal of Quantitative Linguistics* 8(1). 33–50.

Gries, Stefan Th. 2005. Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research* 34(4). 365–399.

Gries, Stefan Th. 2013. Sources of variability relevant to the cognitive sociolinguist, and corpus- as well as psycholinguistic methods and notions to handle them. *Journal of Pragmatics* 52. 5–16.

Gries, Stefan Th. 2014. Coll.analysis 3.5. A script for R to compute perform collostructional analyses.

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

Gries, Stefan Th. & Tobias Bernaisch. 2016. Exploring epicentres empirically: Focus on South Asian Englishes. *English World-Wide* 37(1). 1–25.

Gries, Stefan Th. & Sandra C. Deshors. 2015. EFL and/vs. ESL? A multi-level regression modeling perspective on bridging the paradigm gap. *International Journal of Learner Corpus Research* 1(1). 130–159.

Gries, Stefan Th. & Anatol Stefanowitsch. 2004. Extending collostructional analysis: A corpus-based perspective on alternations. *International Journal of Corpus Linguistics* 9(1). 97–129.

Gries, Stefan Th. & Stefanie Wulff. 2005. Do foreign language learners also have constructions? Evidence from priming, sorting, and corpora. *Annual Review of Cognitive Linguistics* 3. 182–200.

Grimm, Scott & Joan Bresnan. 2009. Spatiotemporal variation in the dative alternation: A study of four corpora of British and American English. Paper presented at the 3rd International Conference on Grammar and Corpora, 22-24 September 2009, Mannheim, Germany.

Gropen, Jess, Steven Pinker, Michelle Hollander, Richard Goldberg & Ronald Wilson. 1989. The learnability and acquisition of the dative alternation in English. *Language* 65(2). 203–257.

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

Haddican, William. 2010. Theme-goal ditransitives and theme passivisation in British English dialects. *Lingua* 120(10). 2424–2443.

Harder, Peter. 2010. *Meaning in mind and society: A functional contribution to the social turn in Cognitive Linguistics*. Berlin/New York: Mouton de Gruyter.

Harrell, Frank E. 2016. *rms: Regression Modeling Strategies* (R package version 5.0-1). Available at: https://CRAN.R-project.org/package=rms.

Harris, John. 1984. Syntactic variation and dialect divergence. *Journal of Linguistics* 20(2). 303–327.

Haspelmath, Martin. 2013. Ditransitive constructions: The verb 'give'. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*, Leipzig: Max Planck Institute for Evolutionary Anthropology. Available at: http://wals.info/chapter/105.

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

Hawkins, John A. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press.

Heller, Benedikt. 2018. *Stability and fluidity in syntactic variation world-wide: The genitive alternation across varieties of English*. Leuven: KU Leuven PhD dissertation.

Hernández, Nuria. 2006. User's Guide to FRED (Freiburg English Dialect Corpus). Available at: http://www.freidok.uni-freiburg.de/volltexte/2489/pdf/Userguide_neu.pdf.

Hickey, Raymond. 2004. Irish English: Phonology. In Bernd Kortmann, Edgar W. Schneider, Kate Burridge, Rajend Mesthrie & Clive Upton (eds.), *A handbook of varieties of English*, Vol. 1: Phonology, 68–97. Amsterdam/Philadelphia: Mouton de Gruyter.

Hickey, Raymond. 2010. The Englishes of Ireland: Emergence and transportation. In Andy Kirkpatrick (ed.), *The Routledge handbook of World Englishes*, 76–95. London/New York: Routledge.

Hilpert, Martin. 2008. The English comparative: Language structure and language use. *English Language and Linguistics* 12(3). 395–417.

Hinrichs, Lars & Benedikt Szmrecsanyi. 2007. Recent changes in the function and frequency of Standard English genitive constructions: A multivariate analysis of tagged corpora. *English Language and Linguistics* 11(3). 437–474.

Hoffmann, Sebastian, Marianne Hundt & Joybrato Mukherjee. 2011. Indian English - an emerging epicentre? A pilot study on light verbs in web-derived corpora of South Asian Englishes. *Anglia* 129(3-4). 258–280.

Holmes, Janet. 1997. Maori and Pakeha English: Some New Zealand social dialect data. *Language in Society* 26(1). 65–101.

Hosmer, David & Stanley Lemeshow. 2000. *Applied logistic regression*. New York: Wiley.

Hothorn, Torsten, Peter Buehlmann, Sandrine Dudoit, Annette Molinaro & Mark Van Der Laan. 2006a. Survival ensembles. *Biostatistics* 7(3). 355–373.

Hothorn, Torsten, Kurt Hornik & Achim Zeileis. 2006b. Unbiased recursive partitioning : A conditional inference framework. *Journal of Computational and Graphical Statistics* 15(3). 651–674.

Hothorn, Torsten & Achim Zeileis. 2015. partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research* 16. 3905–3909. Available at: http://jmlr.org/papers/v16/hothorn15a.html.

Huddleston, Rodney & Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Hundt, Marianne & Geoffrey Leech. 2012. "Small is beautiful": On the value of standard reference corpora for observing recent grammatical change. In Terttu Nevalainen & Elizabeth Closs Traugott (eds.), *The Oxford handbook of the history of English*, 175–188. Oxford: Oxford University Press.

Hundt, Marianne, Gerold Schneider & Elena Seoane. 2016. The use of the *be*-passive in academic Englishes: Local versus global usage in an international language. *Corpora* 11(1). 29–61.

Hundt, Marianne & Benedikt Szmrecsanyi. 2012. Animacy in early New Zealand English. *English World-Wide* 33(3). 241–263.

Janitza, Silke, Carolin Strobl & Anne-Laure Boulesteix. 2013. An AUC-based permutation variable importance measure for random forests. *BMC Bioinformatics* 14(119). 1–11.

Jenset, Gard B., Barbara McGillivray & Michael Rundell. 2017. Keeping the English dative alternation in the family: A quantitative corpus-based study of spoken data. Paper presented at the 9th International Corpus Linguistics Conference, 27 July 2017, University of Birmingham.

Jäschke, Katja & Ingo Plag. 2016. The dative alternation in German-English interlanguage. *Studies in Second Language Acquisition* 38(3). 485–521.

Kachru, Braj B. (ed.). 1982. *The Other tongue: English across cultures*. Urbana: University of Illinois Press.

Kachru, Braj B. 1985. Standards, codification and sociolinguistic realism: The English language in the outer circle. In Randolph Quirk & Henry G. Widdowson (eds.), *English in the world: Teaching and learning the language and literatures*, 11–30. Cambridge: Cambridge University Press.

Kendall, Tyler, Joan Bresnan & Gerard Van Herk. 2011. The dative alternation in African American English: Researching syntactic variation and change across sociolinguistic datasets. *Corpus Linguistics and Linguistic Theory* 7(2). 229–244.

Kirk, Ingrid, Jeffrey L. Kallen, Orla Lowry, Anne Rooney & Margaret Mannion. 2007. The ICE-Ireland Corpus (version 1.2). Available at: http://ice-corpora.net/ice/download.htm.

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

Kortmann, Bernd & Kerstin Lunkenheimer (eds.). 2012. *The Mouton World Atlas of Variation in English*. Berlin/New York: Mouton de Gruyter.

Kortmann, Bernd & Kerstin Lunkenheimer (eds.). 2013. *eWAVE*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available at: http://ewave-atlas.org/.

Kortmann, Bernd, Edgar W. Schneider, Kate Burridge, Raj Mesthrie & Clive Upton (eds.). 2004. *A handbook of varieties of English*. Berlin/New York: Mouton de Gruyter.

Kortmann, Bernd & Clive Upton. 2004. Introduction: Varieties of English in the British Isles. In Bernd Kortmann, Edgar W. Schneider, Kate Burridge, Rajend Mesthrie & Clive Upton (eds.), *A handbook of varieties of English*, 25–33. Berlin/New York: Mouton De Gruyter.

Kristiansen, Gitte & Dirk Geeraerts. 2013. Contexts and usage in Cognitive Sociolinguistics. *Journal of Pragmatics* 52. 1–4.

Kruskal, Joseph & Myron Wish. 1978. *Multidimensional scaling*. Newbury Park/London/New Delhi: SAGE Publications.

Kuhn, Max & Kjell Johnson. 2016. *Applied predictive modeling*. New York: Springer.

Kuhn, Max, With Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael

Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan & Tyler Hunt. 2016. *caret: Classification and regression training* (R package version 6.0-73). Available at: https://CRAN.R-project.org/package=caret.

Labov, William. 1963. The social motivation of a sound change. *Word* 19. 273–309.

Labov, William. 1966. *The social stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.

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

Labov, William. 1972b. Some principles of linguistic methodology. *Language in Society* 1(1). 97–120.

Labov, William. 1982. Building on empirical foundations. In Winfred Lehmann & Yakov Malkiel (eds.), *Perspectives on historical linguistics*, 17–92. Amsterdam/Philadelphia: John Benjamins.

Lalla, Barbara & Jean D'Costa. 1990. *Language in exile: Three hundred years of Jamaican Creole*. Tuscaloosa/London: University of Alabama Press.

Larson, Richard K. 1988. On the double object construction. *Linguistic Inquiry* 19(3). 335–391.

Lavandera, Beatriz. 1978. Where does the sociolinguistic variable stop? *Language in Society* 7(2). 171–183.

Leufkens, Sterre. 2013. The transparency of creoles. *Journal of Pidgin and Creole Languages* 28(2). 323–362.

Levey, Stephen. 2010. The Englishes of Canada. In Andy Kirkpatrick (ed.), *The Routledge handbook of World Englishes*, 113–131. London/New York: Routledge.

Levin, Beth. 1993. *English verb classes and alternations: A preliminary investigation*. Chicago: University of Chicago Press.

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

Levy, Roger. 2014. Using R formulae to test for main effects in the presence of higher-order interactions. Available at: http://arxiv.org/abs/1405.2094.

Lourdes G. Tayao, Maria. 2004. Philippine English: Phonology. In Bernd Kortmann, Edgar W. Schneider, Kate Burridge, Rajend Mesthrie & Clive Upton (eds.), *A handbook of varieties of English*, Vol. 1: Phonology, 1047–1059. Amsterdam/Philadelphia: Mouton de Gruyter.

Lourdes S. Bautista, Ma., Jenifer Loy Lising & Danilo T. Dayag. 2004. The Philippines Corpus. Available at: http://ice-corpora.net/ice/download.htm.

Low, Ee-Ling. 2010. English in Singapore and Malaysia: Differences and similarities. In Andy Kirkpatrick (ed.), *The Routledge handbook of World Englishes*, 229–246. London/New York: Routledge.

MacDonald, Maryellen C. 2013. How language production shapes language form and comprehension. *Frontiers in Psychology* 4(226). 1–16.

MacWhinney, Brian. 1997. Second language acquisition and the Competition Model. In Anette M. B. De Groot & Judith F. Kroll (eds.), *Tutorials in bilingualism: Psycholinguistic perspectives*, 113–142. Mahawa, NJ: Lawrence Erlbaum Associates.

Maechler, Martin, Peter Rousseeuw, Anja Struyf, Mia Hubert & Kurt Hornik. 2016. *cluster: Cluster analysis basics and extensions* (R package version 2.0.5).

Mair, Christian. 2002. Creolisms in an emergent standard: Written English in Jamaica. *English World-Wide* 23(1). 31–58.

Mair, Christian. 2013. The world system of Englishes: Accounting for the transnational importance of mobile and mediated vernaculars. *English World-Wide* 34(3). 253–278.

Malchukov, Andrej, Martin Haspelmath & Bernard Comrie (eds.). 2010. *Studies in ditransitive constructions: A comparative handbook*. Berlin/New York: Mouton de Gruyter.

McArthur, Tom. 1998. *The English languages*. Cambridge: Cambridge University Press.

McDonough, Kim. 2006. Interaction and syntactic priming: English L2 speakers' production of dative constructions. *Studies in Second Language Acquisition* 28(2). 179–207.

McFadden, Thomas. 2002. The rise of the *to*-dative in Middle English. In David Lightfoot (ed.), *Syntactic effects of morphological change*, 107–123. Oxford: Oxford University Press.

Meade, Rocky R. 2001. *Acquisition of Jamaican phonology*. Delft: De Systeem Drukkers.

Melchers, Gunnel & Philip Shaw. 2011. *World Englishes*. London: Hodder Education.

Menard, Scott. 2010. *Logistic regression: From introductory to advanced concepts and applications*. Thousand Oaks/London: SAGE Publications.

Mesthrie, Rajend. to appear. World Englishes, second language acquisition, and language contact. In Markku Filppula, Juhani Klemola & Devyani Sharma (eds.), *The Oxford handbook of world Englishes*, Oxford: Oxford University Press.

Mesthrie, Rajend & Rakesh M. Bhatt. 2008. *World Englishes: The study of new linguistic varieties*. Cambridge: Cambridge University Press.

Michaelis, Susanne Maria, Philippe Maurer, Martin Haspelmath & Magnus Huber (eds.). 2013. *APiCS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available at: http://apics-online.info/.

Milin, Petar, Dagmar Divjak, Strahinja Dimitrijevic & R. Harald Baayen. 2016. Towards cognitively plausible data science in language research. *Cognitive Linguistics* 27(4). 507–526.

Milroy, James & Leslie Milroy. 1993. *Real English: The grammar of English dialects in the British Isles*. London: Longman.

Milroy, Leslie. 1980. *Language and social networks*. Baltimore, MD: University Park Press.

Mufwene, Salikoko. 2001. *The ecology of language evolution*. Cambridge: Cambridge University Press.

Mukherjee, Joybrato. 2005. *English ditransitive verbs: Aspects of theory, description and a usage-based model*. Amsterdam/New York: Rodopi.

Mukherjee, Joybrato. 2007. Steady states in the evolution of New Englishes: Present-day Indian English as an equilibrium. *Journal of English Linguistics* 35(2). 157–187.

Mukherjee, Joybrato. 2010a. Corpus-based insights into verb-complementational innovations in Indian English: Cases of nativised semantico-structural analogy. In Alexandra N. Lenz & Albrecht Plewnia (eds.), *Grammar between norm and variation*, 219–241. Frankfurt am Main: Peter Lang.

Mukherjee, Joybrato. 2010b. The development of the English language in India. In Andy Kirkpatrick (ed.), *The Routledge handbook of World Englishes*, 167–180. London/New York: Routledge.

Mukherjee, Joybrato & Stefan Th. Gries. 2009. Collostructional nativisation in New Englishes: Verb-construction associations in the International Corpus of English. *English World-Wide* 30(1). 27–51.

Mukherjee, Joybrato & Sebastian Hoffmann. 2006. Describing verb-complementational profiles of New Englishes: A pilot study of Indian English. *English World-Wide* 27(2). 147–173.

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

Nathan, Lance. 2006. *On the interpretation of concealed questions*. Cambridge, MA: MIT PhD dissertation.

Nelson, Gerald. 1996. The design of the corpus. In Sidney Greenbaum (ed.), *Comparing English worldwide: The International Corpus of English*, 27–35. Oxford/New York: Clarendon Press/Oxford University Press.

Newman, John & Georgie Columbus. 2010. The ICE-Canada Corpus. Version 1. Available at: http://ice-corpora.net/ice/download.htm.

Nihilani, Paroo, Ni Yibin, Anne Pakir & Vincent Ooi. 2002. The Singapore Corpus. Available at: http://ice-corpora.net/ice/download.htm.

Oehrle, Richard Thomas. 1976. *The grammatical status of the English dative alternation*. Cambridge, MA: MIT PhD dissertation.

Oksanen, Jari, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs & Helene Wagner. 2017. *vegan: Community ecology package* (R package version 2.4-3). Available at: https://CRAN.R-project.org/package=vegan.

Olavarría de Ersson, Eugenia & Philip Shaw. 2003. Verb complementation patterns in Indian Standard English. *English World-Wide* 24(2). 137–161.

Osselton, Noel. 1988. Thematic genitives. In Graham Nixon & John Honey (eds.), *An historic tongue: Studies in English linguistics in memory of Barbara Strang*, 138–144. London: Routledge.

Patrick, Peter. 2004. Jamaican Creole: Morphology and syntax. In Bernd Kortmann, Edgar W. Schneider, Kate Burridge, Rajend Mesthrie & Clive Upton (eds.), *A handbook of varieties of English*, Vol. 2: Morphology and syntax, 407–439. Berlin/New York: Mouton de Gruyter.

Pefianco Martin, Isabel. 2010. Periphery ELT: The politics and practice of teaching English in the Philippines. In Andy Kirkpatrick (ed.), *The Routledge handbook of World Englishes*, 247–264. London/New York: Routledge.

Pefianco Martin, Isabel. 2014. Beyond nativization? Philippine English in Schneider's Dynamic Model. In Sarah Buschfeld, Thomas Hoffmann, Magnus Huber & Alexander Kautzsch (eds.), *The evolution of Englishes: The Dynamic Model and beyond*, 70–85. Amsterdam/Philadelphia: John Benjamins.

Pienemann, Manfred. 1998. *Language processing and second language acquisition: Processability theory*. Amsterdam/Philadelphia: John Benjamins.

Pinheiro, José C. & Douglas M. Bates. 2000. *Mixed-effects models in S and S-PLUS*. New York: Springer.

Poplack, Shana & Sali A. Tagliamonte. 2001. *African American English in the diaspora*. Oxford: Blackwell.

Pütz, Martin, Justyna A. Robinson & Monika Reif. 2014. *Cognitive Sociolinguistics: Social and cultural variation in cognition and language use*. Amsterdam/Philadelphia: John Benjamins.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. Harlow: Longman.

R Core Team. 2016. *R: A language and environment for statistical computing*. Vienna. Available at: http://www.r-project.org/.

Ransom, Evelyn. 1979. Definiteness and animacy constraints on passive and double-object constructions in English. *Glossa* 13. 215–240.

Rappaport Hovav, Malka & Beth Levin. 2008. The English dative alternation: The case for verb sensitivity. *Journal of Linguistics* 44(1). 129–167.

Rickford, John R. 1987. *Dimensions of a creole continuum: History, texts, and linguistic analysis of Guyanese Creole*. Stanford, CA: Stanford University Press.

Rickford, John R. 2014. Situation: Stylistic variation in sociolinguistic corpora and theory. *Language and Linguistics Compass* 8(11). 590–603.

Rohdenburg, Günter. 2002. Processing complexity and the variable use of prepositions in English. In Hubert Cuyckens & Günter Radden (eds.), *Perspectives on prepositions*, 79–100. Tübingen: Niemeyer.

Rosenbach, Anette. 2005. Animacy versus weight as determinant of grammatical variation in English. *Language* 81(3). 613–644.

Rosenfelder, Ingrid, Susanne Jantos, Nicole Höhn & Christian Mair. 2009. Manual for the Jamaican component (ICE-JA). Available at: http://ice-corpora.net/ice/download.htm.

Ross, John Robert. 2004. Nouniness. In Bas Aarts, David Denison, Evelien Keizer & Gergana Popova (eds.), *Fuzzy grammar: A reader*, 351–422. Oxford: Oxford University Press.

Rosseel, Laura. 2017. *New approaches to measuring the social meaning of variation: Exploring the Personalized Implicit Association Test and the Relational Responding Task*. Leuven: KU Leuven PhD dissertation.

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

Sand, Andrea. 2004. Shared morpho-syntactic features in contact varieties of English: Article use. *World Englishes* 23(2). 281–298.

Sankoff, David. 1988. Sociolinguistics and syntactic variation. In Frederick J. Newmeyer (ed.), *Linguistics: The Cambridge survey*, 140–161. Cambridge: Cambridge University Press.

Sankoff, David & William Labov. 1979. On the uses of variable rules. *Language in society* 8(2-3). 189–222.

Sankoff, Gillian, Pierrette Thibault, Naomi Nagy, Hélène Blondeau, Marie-Odile Fonollosa & Lucie Gagnon. 1997. Variation in the use of discourse markers in a language contact situation. *Language Variation and Change* 9(2). 191–217.

Savage, Ceri, Elena Lieven, Anna Theakston & Michael Tomasello. 2003. Testing the abstractness of children's linguistic representations: Lexical and structural priming of syntactic constructions in young children. *Developmental Science* 6(5). 557–567.

Schilk, Marco, Tobias Bernaisch & Joybrato Mukherjee. 2012. Mapping unity and diversity in South Asian English lexicogrammar: Verb-complementational preferences across varieties. In Marianne Hundt & Ulrike Gut (eds.), *Mapping unity and diversity world-wide: Corpus-based studies of New Englishes*, 137–165. Amsterdam: John Benjamins.

Schilk, Marco, Joybrato Mukherjee, Christopher Nam & Sach Mukherjee. 2013. Complementation of ditransitive verbs in South Asian Englishes: A multifactorial analysis. *Corpus Linguistics and Linguistic Theory* 9(2). 187–225.

Schneider, Edgar W. 2003. The dynamics of New Englishes: From identity construction to dialect birth. *Language* 79(2). 233–281.

Schneider, Edgar W. 2007. *Postcolonial English: Varieties around the world*. Cambridge: Cambridge University Press.

Schneider, Edgar W. 2011. *English around the world: An introduction*. Cambridge: Cambridge University Press.

Schneider, Edgar W. 2014. New reflections on the evolutionary dynamics of World Englishes. *World Englishes* 33(1). 9–32.

Scott-Phillips, Thomas C. & Simon Kirby. 2010. Language evolution in the laboratory. *Trends in Cognitive Sciences* 14(9). 411–417.

Setter, Jane, Cathy S. P. Wong & Brian H. S. Chan. 2010. *Hong Kong English*. Edinburgh: Edinburgh University Press.

Shastri, S.V. & Gerhard Leitner. 2002. The ICE India Corpus. Available at: http://ice-corpora.net/ice/download.htm.

Shih, Stephanie & Jason Grafmiller. 2011. Weighing in on end weight. Paper presented at the 85th Annual Meeting of the Linguistic Society of America, 9 January 2011, Pittsburgh, PA.

Siegel, Jeff, Benedikt Szmrecsanyi & Bernd Kortmann. 2014. Measuring analyticity and syntheticity in creoles. *Journal of Pidgin and Creole Languages* 29(1). 49–85.

Siemund, Peter. 2013. *Varieties of English: A typological approach*. Cambridge: Cambridge University Press.

Siewierska, Anna & Willem Hollmann. 2007. Ditransitive clauses in English with special reference to Lancashire dialect. In Mike Hannay & Gerard J. Steen (eds.), *Structural-functional studies in English grammar: In honour of Lachlan Mackenzie*, 83–102. Amsterdam: John Benjamins.

Silva-Corvalán, Carmen. 1986. On the problem of meaning in sociolinguistic studies of syntactic variation. In Dieter Kastovsky & Aleksander Szwedek (eds.), *Linguistics across historical and geographical boundaries*, 111–124. Berlin/New York: Mouton de Gruyter.

Smyth, Ronald, Gary Prideaux & John Hogan. 1979. The effect of context on dative position. *Lingua* 47(1). 27–42.

Stallings, Lynne M. & Maryellen C. MacDonald. 2011. It's not just the "heavy NP": Relative phrase length modulates the production of heavy-NP shift. *Journal of Psycholinguistic Research* 40(3). 177–187.

Stefanowitsch, Anatol & Stefan Th. Gries. 2003. Collostructions : Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2). 209–243.

Stefanowitsch, Anatol & Stefan Th. Gries. 2005. Covarying collexemes. *Corpus Linguistics and Linguistic Theory* 1(1). 1–43.

Stefanowitsch, Anatol & Stefan Th. Gries. 2009. Corpora and grammar. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics: An international handbook*, Vol. 2, 933–951. Berlin/New York: Mouton de Gruyter.

Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin & Achim Zeileis. 2008. Conditional variable importance for random forests. *BMC Bioinformatics* 9(307). Available at: http://www.biomedcentral.com/1471-2105/9/307.

Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis & Torsten Hothorn. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* 8(25). Available at: http://www.biomedcentral.com/1471-2105/8/25.

Szmrecsanyi, Benedikt. 2004. On operationalizing syntactic complexity. In *Le poids des mots. Proceedings of the 7th International Conference on Textual Data Statistical Analysis*, Vol. 2, 1032–1039. Louvain-la-Neuve: Presses universitaires de Louvain.

Szmrecsanyi, Benedikt. 2005. Language users as creatures of habit: A corpus-based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory* 1(1). 113–150.

Szmrecsanyi, Benedikt. 2008. Corpus-based dialectometry: Aggregate morphosyntactic variability in British English dialects. *International Journal of Humanities and Arts Computing* 2(1-2). 279–296.

Szmrecsanyi, Benedikt. 2013. *Grammatical variation in British English dialects: A study in corpus-based dialectometry*. Cambridge: Cambridge University Press.

Szmrecsanyi, Benedikt. 2017a. The register-specificity of variation grammars. Paper presented at the 39th Annual meeting of the German Linguistic Society, 8-10 March 2017, Saarbrücken, Germany.

Szmrecsanyi, Benedikt. 2017b. Variationist sociolinguistics and corpus-based variationist linguistics: Overlap and cross-pollination potential. *Canadian Journal of Linguistics/Revue Canadienne De Linguistique* 42(4). 685–701.

Szmrecsanyi, Benedikt, Jason Grafmiller, Joan Bresnan, Anette Rosenbach, Sali A. Tagliamonte & Simon Todd. 2017. Spoken syntax in a comparative perspective: The dative and genitive alternation in varieties of English. *Glossa* 2(1). 1–17.

Szmrecsanyi, Benedikt, Jason Grafmiller, Benedikt Heller & Melanie Röthlisberger. 2016. Around the world in three alternations: Modeling syntactic variation in varieties of English. *English World-Wide* 37(2). 109–137.

Szmrecsanyi, Benedikt & Lars Hinrichs. 2008. Probabilistic determinants of genitive variation in spoken and written English: A multivariate analysis across time, space, and genres. In Terttu Nevalainen, Irma Taavitsainen, Päivi Pahta & Minna Korhonen (eds.), *The dynamics of linguistic variation: Corpus evidence on English past and present*, 291–309. Amsterdam/Philadelphia: John Benjamins.

Szmrecsanyi, Benedikt & Bernd Kortmann. 2009. The morphosyntax of varieties of English worldwide: A quantitative perspective. *Lingua* 119(11). 1643–1663.

Szmrecsanyi, Benedikt & Melanie Röthlisberger. to appear. World Englishes from the perspective of dialect typology. In Daniel Schreier, Marianne Hundt & Edgar W. Schneider (eds.), *The Cambridge handbook of World Englishes*, Cambridge: Cambridge University Press.

Tagliamonte, Sali A. 2002. Comparative sociolinguistics. In J. K. Chambers, Peter Trudgill & Natalie Schilling-Estes (eds.), *The handbook of language variation and change*, 729–763. Malden/Oxford: Blackwell.

Tagliamonte, Sali A. 2006. *Analysing sociolinguistic variation*. Cambridge: Cambridge University Press.

Tagliamonte, Sali A. 2012. *Variationist sociolinguistics: Change, observation, interpretation*. Malden, MA: Wiley-Blackwell.

Tagliamonte, Sali A. 2014. A comparative sociolinguistic analysis of the dative alternation. In Rena Torres-Cacoullos, Nathalie Dion & André Lapierre (eds.), *Linguistic variation: Confronting fact and theory*, 297–318. London/New York: Routledge.

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

Tagliamonte, Sali A., Mercedes Durham & Jennifer Smith. 2014. Grammaticalization at an early stage: Future *be going to* in conservative British dialects. *English Language and Linguistics* 18(1). 75–108.

Tagliamonte, Sali A. & Jennifer Smith. 2002. "Either it isn't or it's not": NEG/AUX contraction in British dialects. *English World-Wide* 2(23). 251–281.

Tagliamonte, Sali A. & Jennifer Smith. 2005. No momentary fancy! The zero in English dialects. *English Language and Linguistics* 9(2). 289–309.

Theijssen, Daphne. 2012. *Making choices: Modelling the English dative alternation*. Nijmegen: Radboud Universiteit PhD dissertation.

Theijssen, Daphne, Joan Bresnan, Marilyn Ford & Lou Boves. 2011. In a land far far away... A probabilistic account of the dative alternation in British, American, and Australian English. Manuscript.

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

Thompson, Sandra A. 1990. Information flow and dative shift in English discourse. In Jerold A. Edmondson, Crawford Feagin & Peter Mühlhäusler (eds.), *Development and diversity: Language variation across time and space. A Festschrift for Charles-James N. Bailey*, 239–253. Arlington: SIL and University of Texas.

Tomasello, Michael. 2003. *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Trudgill, Peter. 1974. *The social differentiation of English in Norwich*. Cambridge: Cambridge University Press.

Trudgill, Peter. 1984. *Language in the British Isles*. Cambridge: Cambridge University Press.

van den Bosch, Antal & Joan Bresnan. 2015. Modeling dative alternations of individual children. In *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning*, 103–112. Lisbon, Portugal: Association for Computational Linguistics.

Venables, William N. & Brian D. Ripley. 2002. *Modern applied statistics with S*. New York: Springer. Available at: http://www.stats.ox.ac.uk/pub/MASS4.

Vine, Ingrid, Janet Holmes & Laurie Bauer. 1999. Guide to the New Zealand component of the International Corpus of English (ICE-NZ). Available at: http://ice-corpora.net/ice/download.htm.

Walker, James A. & Miriam Meyerhoff. 2013. Studies of the community and the individual. In Robert Bayley, Richard Cameron & Ceil Lucas (eds.), *The Oxford handbook of Sociolinguistics*, 175–194. Oxford: Oxford University Press.

Wasow, Thomas. 1997a. End-weight from the speaker's perspective. *Journal of Psycholinguistic Research* 26(3). 347–361.

Wasow, Thomas. 1997b. Remarks on grammatical weight. *Language Variation and Change* 9(1). 81–105.

Wasow, Thomas. 2002. *Postverbal Behavior*. Stanford: CSLI Publications.

Wasow, Thomas & Jennifer Arnold. 2003. Post-verbal constituent ordering in English. In Günter Rohdenburg & Britta Mondorf (eds.), *Determinants of grammatical variation in English*, 119–154. Amsterdam: Mouton de Gruyter.

Wasow, Thomas & Jennifer Arnold. 2005. Intuitions in linguistic argumentation. *Lingua* 115(11). 1481–1496.

Wee, Lionel. 2014. The evolution of Singlish in late modernity: Beyond phase 5? In Sarah Buschfeld, Thomas Hoffmann, Magnus Huber & Alexander Kautzsch (eds.), *The evolution of Englishes: The Dynamic Model and beyond*, 126–141. Amsterdam/Philadelphia: John Benjamins.

Weiner, Judith & William Labov. 1983. Constraints on the agentless passive. *Journal of Linguistics* 19(1). 29–58.

Weinreich, Uriel, William Labov & Marvin Herzog. 1968. *Empirical foundations for a theory of language change*. Austin: University of Texas Press.

Williams, Robert S. 1994. A statistical analysis of English double object alternation. *Issues in Applied Linguistics* 5(1). 37–58.

Wolfe-Quintero, Kate. 1993. The dative alternation in English. *University of Hawai'i Working Papers in ESL* 11(2). 91–120.

Wolk, Christoph, Joan Bresnan, Anette Rosenbach & Benedikt Szmrecsanyi. 2013. Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica* 30(3). 382–419.

Yáñez-Bouza, Nuria & David Denison. 2015. Which comes first in the double object construction? *English Language and Linguistics* 19(2). 247–268.

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

Zehentner, Eva. 2016. *On competition and cooperation in Middle English ditransitives*. Wien: Universität Wien PhD dissertation.

Zehentner, Eva. 2017. Ditransitives in Middle English: On semantic specialisation and the rise of the dative alternation. *English Language and Linguistics* 1–27.

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


1425