1 Introduction

This book will establish a canon of state-of-the-art quantitative methods to responsibly and rigorously analyze variationist datasets from a comparative perspective. In this spirit, we will showcase various theoretically exciting intersections between variationist linguistics and related subfields, including dialectology and dialect typology, comparative linguistics, probabilistic linguistics, usage-based theoretical linguistics, psycholinguistics, and research on English as a world language.

As a case study, we will distill key findings and methodological innovations from a five-year research project entitled "Exploring probabilistic grammar(s) in varieties of English around the world" about the scope and limits of grammatical variation in a global language such as English. In this book, we adopt the variationist methodology and take a particular interest in how people choose between "alternate ways of saying 'the same' thing" (Labov, 1972, 188). In so doing, the book breaks new ground by marrying the spirit of Probabilistic Grammar research (which posits that grammatical knowledge is experience-based and partially probabilistic – see Grafmiller et al., 2018) to research along the lines of the English worldwide paradigm (which is concerned with the dialectology and sociolinguistics of postcolonial English-speaking communities around the world – see Schneider, 2007). The overarching objective, then, is to understand the plasticity of probabilistic knowledge of English grammar, on the part of language users with diverse regional and cultural backgrounds: how different are the ways a speaker of, say, British English chooses between different ways of saying the same thing (e.g. *look up the word* vs. *look the word up*) from how a speaker of, say, Canadian English chooses? To address this question, we investigate three grammatical alternations (see (1) to (3)) in some nine varieties of English around the world (British English, Canadian English, Irish English, New Zealand, English, Hong Kong English, Indian English, Jamaican English, Philippines English, and Singapore English).

https://doi.org/10.1017/9781108863742.001 Published online by Cambridge University Press

(1) **The genitive alternation in English:**

a. Two other journalists who wrote a book criticising [the president]$_{possessor}$'s [brother]$_{possessum}$ were ordered to pay £6.3 million in fines (GloWbE AU B vexnews.com)
(the *s*-genitive)

b. Can you imagine a couple of years after WW2 the allies permitting [the brother]$_{possessum}$ of [the president]$_{possessor}$ bankrupting the central bank through embezzlement and getting away with it? (GloWbE GB G guardian.co.uk)
(the *of*-genitive)

(2) **The dative alternation in English:**

a. A victim will be asked to give$_{verb}$ [the police]$_{recipient}$ [a statement]$_{theme}$ explaining what has happened. (GloWbE CA G slsedmonton.com)
(the ditransitive dative)

b. Neither of them gave$_{verb}$ [a statement]$_{theme}$ to [the police]$_{recipient}$. (GloWbE JM G jamaicaobserver.com)
(the prepositional dative)

(3) **The particle placement alternation in English:**

a. For all my second language readers: no need to look$_{verb}$ [the word]$_{NP}$ up$_{particle}$ in the dictionary … (GloWbE NZ B dedepuppets.com)
(the split variant)

b. Look$_{verb}$ up$_{particle}$ [the word]$_{NP}$ in a dictionary and write down its meaning in a vocabulary notebook. (GloWbE US G artofmanliness.com)
(the continuous variant)

The analysis is mostly based upon observational corpus analysis but will be supplemented behaviorally by rating task experiments. Why do we need these two sources of evidence? On a practical level, corpora provide potentially massive amounts of data that can be analyzed at comparatively low cost (especially if the analysis relies on pre-existing corpora, as we do in this book). More substantially speaking though, corpora cover naturalistic language usage, not behavior in (more or less) artificial experiments, which is why corpus findings are ecologically valid in a way that experimental findings are not (see Campbell-Kibler, 2010 for discussion). On the other hand, experiments can directly target specific phenomena, variables, and constraints in a way that corpus analysis cannot. What is more, rating task experiments (the type of experiment we will be relying on) in particular explore metalinguistic judgments, a facet of linguistic competence that is not covered by corpora (which

https://doi.org/10.1017/9781108863742.001 Published online by Cambridge University Press

cover production and to some extent comprehension). Finally, it is always a good idea to strive for methodological pluralism (see Klavan and Divjak, 2016).

Thus we aim to sketch a picture of probabilistic grammar variation across different native and nonnative varieties of English, and to develop a method for exploring indigenization patterns which builds upon established methods in comparative sociolinguistics while expanding our analytical toolkit to include methods common in dialectology and in psycholinguistics. The specific research questions that will guide our inquiry include the following: For a given alternation, how consistent are the probabilistic effects of the variable grammar's constraints across varieties? Do some alternations vary more than others with respect to their probabilistic conditioning? Are there some (types of) constraints that are more variable than others? How and where to draw boundaries between distinct probabilistic variable grammars? To what extent can the patterns we observe in corpus data be replicated in rating task experiments? Do the crossvarietal patterns we find align with our current understanding of typological variation among varieties of English?

In a nutshell, we may summarize the key findings to be discussed at length in the remainder of this book as follows. Probabilistic grammars across World Englishes are overall surprisingly stable: on a scale between 0 and 1, where 0 indicates total dissimilarity and 1 indicated total identity, the overall similarity of the alternation phenomena under study calculates as approximately 0.7. Effect directions are stable across varieties. If a particular constraint favors a particular grammatical outcome in a given variety, it will also do so in the other varieties. In contrast, strength of effects vary. For example, animacy may have strong effects on grammatical outcomes in variety A, but comparatively weaker effects in variety B . We will also see that different alternations are differentially hospitable to what we call probabilistic indigenization: for example, the particle placement alternation is (probably in function of its comparatively strong lexical anchoring) particularly malleable. On the interpretational plane, we often see a dialect-typological split between Inner Circle (ENL) and outer Circle (ESL) varieties. Finally, experiments and corpus analysis converge largely but not entirely.

We note that variation studies of the kind presented in this book represent an increasingly popular area in linguistics – they are becoming increasingly entrenched in curriculum design; variation conferences are becoming ever larger and more numerous; and linguists more broadly are increasingly engaging with variation (see Nagy and Hoffman, 2017 for discussion). But this growing interest also means that the field is in danger of fragmenting into different research communities with different foci that do not necessarily talk as much to each other as they should. Against this backdrop, one of the aims of the book is to cross-pollinate different research tracks in variation studies, in one monograph with a coherent empirical focus. On the theoretical plane, the

book prescribes a "balanced diet" (Guy, 2014, 65) to model and interpret variation as the association of conventional rules or constraints with probabilities learned from experience.

In what follows, we briefly discuss some key concepts and research orientations that take center stage in the remaining chapters.

1.1 Variationist Sociolinguistics and Corpus-Based Variationist Linguistics

This book is an exercise in variation analysis. Specifically, we use the variationist method to study variation between grammatical variants that are in principle available to all members of the speech communities under study. The variationist method is designed to investigate quantitatively how speakers choose between "alternate ways of saying 'the same' thing" (Labov, 1972, 188) as a function of properties of the linguistic contexts and of language-external factors.

The variationist method is the cornerstone of the field of Variationist Sociolinguistics, also known as the Language Variation and Change (LVC) paradigm. Variationist Sociolinguistics is a research orientation in sociolinguistics pioneered by William Labov in the 1950s and 1960s (see e.g. Labov, 1963, 1966) dedicated to the rigorous and quantitative study of the interaction between linguistic variation and linguistic change based, typically, on observational data (for instance, corpora covering sociolinguistic interviews). Most work in variationist sociolinguists models the way language users choose between different ways of expressing the same meaning or grammatical function subject to both language-internal constraints (that is, properties of the linguistic context) and language-external constraints (such as age, gender, register, or geography). A key concept in variationist sociolinguistics is that of the linguistic variable (i.e. a particular meaning or function the expression of which is variable) and linguistic variants (particular forms which come under the remit of a particular variable).

Over the last few decades, the variationist methodology has also become popular outside of Variationist Sociolinguistics proper. In particular, in corpus linguistics a new subfield has emerged that Szmrecsanyi (2017) calls corpus-based variationist linguistics (or CVL for short). Compared to other methodologies in corpus linguistics, the focus in CVL is on the conditioning of variation, and not so much on text frequencies. Compared to Variationist Sociolinguistics, CVL analysts tend to be more interested in language-internal constraints on variation than in language-external factors. Also, CVL analysts are more enthusiastic than variationist sociolinguists to consider other registers beside vernacular speech. What both CVL and variationist sociolinguistics share in common is that both orientations carefully define variables

and variants and follow the Principle of Accountability (Labov, 1969, 738) to understand the conditioning of variation.

1.2 Comparative Linguistics and Comparative Variation Analysis

Because we will be interested in the extent to which regional varieties of English are different or equivalents in terms of how language users make grammatical choices, this book also comes under the remit of comparative linguistics. Assessing the similarity or dissimilarity of language systems across varieties, dialects, or languages for that matter is an important topic of theoretical significance in comparative linguistics, including in crosslinguistic typology, dialectology, and sociolinguistics. There is sure enough a rich literature on how to assess such similarity. Much of this literature, however, is based on fairly decontextualized data that are more about competence than about usage, such as reference grammars, dialect atlases (such as e.g. The Survey of English Dialects; Orton and Dieth, 1962), or crosslinguistic surveys (such as e.g. The World Atlas of Language Structures; Dryer and Haspelmath, 2013). Data sources like these are valuable, but not of much use in variationist linguistics: they typically cover the inventory of forms and variants, but do not provide information about probabilistic variation patterns.

Therefore, we will use comparative methods that are designed to investigate usage data about probabilistic variation patterns. We draw inspiration from two traditions of comparative variation analysis. The first tradition goes back to seminal work by Shana Poplack and Sali Tagliamonte (Poplack and Tagliamonte, 1989) and is now a subfield in variationist sociolinguistics known as *comparative sociolinguistics* (see Tagliamonte, 2001). The name of the game in comparative sociolinguistics is to investigate the conditioning of variation in a small number of varieties or dialects for the sake of determining if these varieties or dialects are historically related.

The second comparative tradition that will inform the variation analyses presented here has been developed in corpus-based variationist linguistics (see Section 1.1) and builds on foundational work carried out by Joan Bresnan, Jennifer Hay, Lars Hinrichs and others in the 2000s (see Hinrichs and Szmrecsanyi, 2007; Bresnan and Hay, 2008). Here, the focus is not necessarily on historical relatedness – instead, the research questions addressed in this line of work are the following: are varieties diverging or converging? What are the constraints that are particularly stable or unstable across varieties? What can differences and similarities across varieties tell us about the nature of knowledge that language have about probabilistic grammars? Of note, comparative variation analysis in this spirit can be backed up by rating task experiments (Bresnan and Ford, 2010), and this is exactly what we are going to do in this book as well.

https://doi.org/10.1017/9781108863742.001 Published online by Cambridge University Press

1.3 Dialectology, Dialectometry, and Dialect Typology

Dialectology, dialectometry, and dialect typology are to some extent also comparative endeavors, and this book draws inspiration from all three (sub)fields.

Dialectology is concerned with the study of regional varieties of language and has a long history that goes back at least to the nineteenth century (see Chambers and Trudgill, 1998). This book is inspired by work in dialectology in that what will take center stage is regional varieties of English. That said, we hasten to add that while traditional dialectology focusses on primarily phonological or lexical features of rural dialects spoken by nonmobile old rural males (NORMs), we will be interested in the grammar of more acrolectal international standard varieties of English spoken and written by all kinds of language users. We also deviate from standard practice in traditional dialectology in that our empirical analysis is not based on questionnaires or survey data (which are the customary datasources in traditional dialectology – see Anderwald and Szmrecsanyi, 2009), but on corpora and experiments.

Dialect typology (also known as sociolinguistic typology) is a subfield in dialectology that explores the intersection between dialectology and typology. Typologists seek to categorize human languages based on their structural differences and similarities, and similarly dialect typologists take an interest in categorizing dialects and varieties of the same language. Sometimes dialect typology is used interchangeably with sociolinguistic typology, and in sociolinguistic typology in particular there is an interest in the "extent to which differences of linguistic structure, whether within or between languages, can be ascribed to or explained in terms of features of the society in which the dialects in question are spoken" (Trudgill, 1996, 3; see also Trudgill, 2011; Röthlisberger and Szmrecsanyi, 2019). Dialect typology will play a role in this book because we will systematically distinguish between native L1 varieties of English (such as Canadian English and New Zealand English) and indigenized L2 varieties of English (such as Indian English or Hong Kong English).

Dialecto*metry* is a subfield in dialectology that specializes in measuring, visualising, and analysing aggregate dialect similarities or distances as a function of properties of geographic space (seminal work includes Séguy, 1971, Goebl, 1982, and Nerbonne et al., 1999). Thus, whereas traditional dialectologists study in depth a typically small number of features deemed interesting in a typically correspondingly small number of dialects, dialectometricians explore relationships between a large amount of dialect locations based on a large amount of features. In this endeavor, dialectometrical analysis strongly relies on quantification, cartographic visualisation and exploratory data analysis for the sake of inferring patterns from feature aggregates. This book draws inspiration from dialectometry in that we study multiple grammatical alternations in multiple varieties of English, with one of our aims being the discovery of general patterns in a bird's eye perspective.

1.4 Probabilistic Linguistics and Probabilistic Grammar

Probabilistic Linguistics is a research orientation whose point of departure is that probabilistic patterns and gradience have been shown to be pervasive on all levels of language. Given this pervasiveness, Probabilistic Linguistics seeks to complement more traditional structural/categorical/generative theorizing by exploring the extent to which gradient rules, to be discovered through quantitative modeling using the mathematics of uncertainty, can predict (aspects of) linguistic knowledge and of linguistic usage (see the papers in Bod et al., 2003 for more discussion and case studies). It is clear that quantitative variationist (socio)linguistics (see Section 1.1) is essentially a variety of Probabilistic Linguistics. Variationist sociolinguists, after all, have been busy analyzing variation patterns probabilistically for decades (see e.g. Cedergren and Sankoff, 1974).

Needless to say, this book is (among other things) engaging in Probabilistic Linguistics. And even more specifically, this book will engage in (comparative) Probabilistic Grammar analysis (see Grafmiller et al., 2018, which is summarized in the following discussion). As our literature review in Chapter 2 will show, it is amply documented that (morpho)syntactic variation within and across varieties of the same language is very systematic, and that the determinants of this variation are numerous, multifactorial, and probabilistic in nature. Against this backdrop, this book endeavors to systematically asses the scope and limits of differentiation between probabilistic grammars in a world language such as English. Crucially, this includes an experimentalist inquiry into the extent to which language users' *knowledge* about probabilistic grammars differs as a function of geography and/or regional identity.

Now, current theorizing about the nature of grammatical knowledge is often trapped in an opposition between fully usage-based (i.e. exemplar-based) approaches (e.g. Pierrehumbert, 2006) and fully rule-based approaches (e.g. Chomsky and Halle, 1968). We do not think that this dichotomy is productive. We are certainly committed to the usage-based notion that grammar is the "cognitive organization of one's experience with language" (Bybee, 2006) – most probabilistic approaches to analyzing variation are actually or inherently usage-based, in that they capitalize on statistical regularities likely derived from experience, yet they associate these quantitative patterns not (only) with surface forms or lexical items (as in pure exemplar models), but with abstract features or constraints. But beyond this commitment, we submit that a hybrid model is necessary to account for variation in all its complexities (see Guy, 2014 for discussion). Of note, both usage- and rule-based models of grammar are mentalistic, in that they view language as a cognitive object, and this common ground is shared by most hybrid models.

The work we report is especially inspired by the variation-centered, usage- and experience-based Probabilistic Grammar approach developed by Joan

https://doi.org/10.1017/9781108863742.001 Published online by Cambridge University Press

Bresnan and collaborators (e.g. Bresnan, 2007 and follow-up work reviewed in Chapter 2). This work, and our work, makes two key assumptions (see Grafmiller et al., 2018, 2–3): First, grammatical knowledge is partially probabilistic in nature, and language users have demonstrably powerful predictive capabilities. Second, this probabilistic knowledge is acquired through language experience, and so is subtly, but dynamically (re)constructed throughout speakers' lives. Needless to say, regional differentiation of the type that takes center stage here is predicted. Note also that this approach is hybrid in nature, as it assumes that conventional rules or constraints are associated with probabilities learned from experience. In other words, the approach we will adopt follows a "balanced diet" (Guy, 2014, 65) by modeling syntactic variation drawing on both qualitative and quantitative aspects.

1.5 Psycholinguistics

Psycholinguistics is a discipline situated at the intersection between psychology and linguistics. Psycholinguists investigate the psychological processes that enable language users to produce, comprehend, process, and acquire language, and the way in which language users store knowledge about language (see e.g. Kennison and Messer, 2014).

This book builds on key insights and makes use of methodologies from psycholinguistics in a number of ways. For one thing, we will be interested in the probabilistic nature of (knowledge of) grammar, an issue that is needless to say inherently relevant to how language is produced and processed, and to how knowledge about language is stored.

Secondly, a number of language-internal constraints on variation that we consider in this book relate to language processing. Consider, for example, priming effects, or surprisal effects. Priming is about the tendency that language users have a preference for reusing syntactic patterns that they have produced or have been exposed to in previous discourse (see e.g. Gries, 2005). Surprisal is about the extent to which material in different slots of constructions tends to co-occur, and about the consequences that these co-occurrence preferences have for syntactic placement preferences (see e.g. Levy and Jaeger, 2007).

Third, we will supplement corpus-based analysis (the customary methodology in variationist (socio)linguistics) with rating task experiments to spot-check the psychological plausibility of our findings. The rating task experiments that we will conduct are inspired by Bresnan (2007, 76–84). That study used a scalar rating task based on corpus materials as stimuli to model subjects' responses regarding the naturalness of syntactic variants in context. Responses were compared to the predictions of a parallel regression model fitted on corpus data. Analysis showed that subjects' gradient naturalness ratings correlated with corpus-generated probabilities. This demonstrates that language users'

implicit knowledge about language must be to some extent probabilistic in nature.

Corpora are an observational data source that covers language production as well as, to some extent, language comprehension (because in naturalistic settings, whatever is spoken/written is designed to be also comprehended). Rating task experiments, on the other hand, tap into subjects' intuitions about the naturalness of grammatical variants given a real-life context. We will thus cover language production, language comprehension, as well as the predictive capacities of language users. The methodological diversity that this book aims for is ultimately motivated by the quest for ecologically valid paradigms (see Klavan and Divjak, 2016 for discussion). In short, we fully agree with (Dąbrowska, 2016a, 488):

> Corpus analysis is absolutely vital to usage-based approaches . . . . In the end, however, corpora can only provide information about frequency of items and frequency of co-occurrence of items. If we want to make claims about speakers' mental representations, corpus data needs to be complemented with experimental research.

1.6 English as a World Language

In this book, we cover variation in multiple varieties of English around the world. Thanks to a history of colonial expansion and other favorable socio-historical circumstances, English is fairly unique in having diversified into a language with a wide range of postcolonial varieties of English (a.k.a. "New Englishes") around the world. This diversity is particularly interesting from the point of view of dialect typology (see Section 1.3), and so we cover both native mother-tongue, "Inner Circle" (Kachru, 1992) varieties (e.g. New Zealand English) as well as non-native indigenized second-language "Outer Circle" (Kachru, 1992) varieties (e.g. Hong Kong English).

Research on the scope and limits of phonological and grammatical variation within and across varieties of English around the world has in recent years engendered a lively research field. Key results of research on the "English language complex" (see McArthur, 2003, 56; Mesthrie and Bhatt, 2008, 1–3) include the finding that the structural make-up of postcolonial varieties of English can be predicted by the particular communicative needs of the colonizers and the colonized (Schneider, 2007), and that there is a "World System of Englishes" in which differential associations with prestige shape the hierarchical structure of, and relationships between, World Englishes (Mair, 2013).

Variationist research activity on World Englishes is clearly picking up (see Chapter 2 for a review), but a shortcoming of much previous research on the English language complex is an often primarily descriptive interest in the variable presence or absence of particular features in particular varieties, or in usage frequencies of grammatical patterns. But while feature inventories and

usage frequencies are no doubt interesting, they do not necessary address the most interesting part of the story, which is: Do language users' probabilistic grammars differ across varieties of English, and if so, to what extent? This is the central question that we are going to address in this book.

1.7 Structure

The remainder of this book is structured as follows:

Chapter 2 surveys the literature on variation in general and on grammatical variables (a.k.a. "alternations") in particular. Next, we review well-known grammatical variables/alternations in English as well as previous comparative investigations of grammatical alternations in English. Last but not least, we discuss in detail previous variationist work on the three alternations subject to study here: the genitive alternation, the dative alternation, and the particle placement alternation.

Chapter 3 begins with a review of the World Englishes and dialect typology literature. Next, we introduce the nine regional varieties of English under study in the book with a brief summary of relevant aspects of their sociohistories and linguistic profiles: British English, Canadian English, Irish English, New Zealand English, Hong Kong English, Indian English, Jamaican English, Philippines English, and Singapore English. These varieties are a fairly representative sample covering both native (or "Inner Circle" – see Kachru, 1992) varieties and non-native (or "Outer Circle") varieties.

Chapter 4 kicks off with a general discussion of the common data types used in variationist linguistics. Next, we present the primary data sources we use in the study. To study variation in production, we tap into corpus data from the International Corpus of English (ICE) (Greenbaum, 1991) and the Corpus of Global Web-based English (GloWbE) (Davies and Fuchs, 2015). After introducing the corpora, we describe the procedures for identifying and extracting interchangeable tokens of each alternation, and detail the annotation procedures for a wide range of constraints including, for example, the principle of end weight (longer constituents tend follow shorter constituents) and animacy effects (animate constituents tend to occur early).

Chapter 5 interrogates corpus data to analyze the three alternations subject to study one-by-one using a battery of state-of-the art analysis techniques, including – in addition to customary descriptive statistics – conditional random forest modeling and mixed-effects logistic regression analysis. The goal of the chapter is to uncover qualitative generalizations: for example, we see that while effect directions of constraints on variation are generally stable across varieties of English, effects strengths can be significantly different.

Chapter 6 is inspired by work in comparative sociolinguistics and quantitative dialectometry. We use a corpus-based method (Variation-Based Distance and Similarity Modeling – VADIS for short) to quantify the similarity between, and coherence across, the varieties of English under study as a function of the correspondence of the ways in which language users choose between different ways of saying the same thing. Key findings include the result that probabilistic grammars are remarkably stable across varieties, but that coherence across alternations is not perfect.

Chapter 7 examines the extent to which contrasts uncovered in the corpus analyses in Chapters 5 and 6 can be replicated in an experimental acceptability judgment task. To compare ratings to corpus model predictions, we use a variant preference rating task modeled after the work of Bresnan and Ford (2010) in which participants rate the naturalness of alternative syntactic forms by distributing points between two alternatives. This experimental paradigm is relatively new but increasingly popular. The hypothesis is that the ratings suggested by participants (who are provided with the surrounding context of the given corpus example) correlate significantly with the probabilities predicted by the corpus models. The results are in line with our expectation – the splits people suggest are typically in line with the splits predicted by corpus-based regression models. Our results provide further evidence that linguistic choices in both production and comprehension are sensitive to the quantitative distributions of various contextual cues, that is, that grammatical knowledge is to some extent probabilistic. However, there remain some subtle discrepancies between our ratings data and corpus models that raise important questions about the comparability of different observational and experimental methods. We consider a number of these questions in the discussion.

Chapter 8 summarizes the study's key findings, and discusses these findings against the backdrop of the various frameworks to which the book is relevant, including World Englishes research and (Labovian) variationist sociolinguistics, but also, for example, dialectometry (e.g. Nerbonne et al., 1999) and general usage, end experience-based linguistics (e.g. Bybee, 2010). We also highlight the application potential of the methodological innovations presented in the book, and conclude with some general reflection on where the road ahead may lead.