CHAPTER 1

Introduction

This book aims to describe the fundamental dimensions of linguistic variation in English worldwide and to analyze the respective role of variety and register in structuring such variation. To accomplish this task, a wide-angle perspective is adopted that looks simultaneously at many individual features (precisely: 236) in a large number of corpus texts (7,309) to quantify their linguistic similarity and difference. Empirically, the study draws on extensive data from the International Corpus of English (ICE) project as well as user-generated content from the social networking service Twitter. In terms of method, the multidimensional (MD) technique of factor analysis (Biber 1988) provides the statistical tool of choice, which offers a good balance between quantitative robustness and descriptive nuance. The results corroborate findings from earlier, more locally restricted MD studies on a global level and challenge the importance of geography in current World Englishes research.

1.1 Research Context

The analysis presented in this book is situated at a point of convergence between several fields of linguistic inquiry: research in the World Englishes paradigm,[1] the growing field of aggregation-based analysis of linguistic variation (e.g. Szmrecsanyi & Wälchli 2014), and the MD tradition of corpus-based register analysis (Biber & Conrad 2009; Biber 1988). These approaches have much in common, such as their recognition of variation as an essential property of human language and their insistence on usage-based, empirical data as the fundament of theoretical model-building.

[1] A number of alternative labels exist, most of which are indebted to a particular theoretical perspective, e.g.: *the English language complex*, *English as a global language*, *English world-wide* (see Seargeant 2010). I use the terms English worldwide and World Englishes as notationally equivalent, general designators, without intending to take a theoretical stance.

Nonetheless, each of them places the focus on different patterns and extralinguistic explanations of variation. Combining them within a common analytical framework is a novel approach that requires some stocktaking of histories, basic assumptions and differences in analytical focus.

1.1.1 *World Englishes*

World Englishes as an independent field of research starts from the simple observation that, after a long process of British colonial expansion followed by decolonization and linguistic fragmentation, a monolithic notion of Standard English is no longer theoretically adequate. The consequence has been a subdivision of standard Englishes in a quasi-dialectological framework, as is well reflected in the subtitle of an early publication (Trudgill & Hannah 1982), which mentions "varieties of Standard English." The extralinguistic category along which such varieties are stratified is usually the political-territorial unit of the nation-state. The overarching theoretical models in the World Englishes literature, to be presented in more detail in Section 2.1, all take the nation-state as their fundamental structuring category.

With the advent of corpus linguistics, and specifically the compilation of the ICE corpora (Greenbaum & Nelson 1996; Greenbaum 1991), close structural comparisons of different situational registers and increasingly also demographic factors across varieties has become a feasible and productive enterprise (see, e.g. Hundt & Gut 2012). Such research continues to be instrumental in further refining descriptive detail in global English studies. It does so by taking a decidedly usage-based approach, developing generalizations directly from samples of authentic, situated language, and by looking not only at variation *across* national standard varieties, but also *within* each territorially defined speech community. For instance, Fuchs and Gut (2015) study demographic and stylistic variation within Nigerian English in relation to the progressive aspect, and Davydova (2019) analyzes internal, sociolinguistic differentiation in the quotative system of different varieties of English. The process of descriptive refinement is ongoing and supported by an elaboration of the corpus database in terms of quantitative (e.g. Davies & Fuchs 2015) and historical (e.g. Collins 2015a) scope.

These developments notwithstanding, much of current theorizing in World Englishes takes a relatively "top-down" perspective, categorizing varieties along preconceived national-historical lines rather than developing their relationships in linguistic terms first. Authors who reject such

categorizations (e.g. Pennycook 2007) offer valid critiques, but no principled system to relate linguistic variation in different contexts. Hundt (2009a) provides a terminology for different kinds of development in postcolonial Englishes. However, while her account adds systematicity to the description it still concerns isolated features and offers no formalized way of fusing individual cases of variation into a synthesized picture of inter-varietal similarity and difference.

1.1.2 Aggregation-Based Linguistics

Such a comprehensive perspective is the express research aim in feature-aggregation-based methods, such as *quantitative dialectology*. Bringing exploratory statistical tools and geographic data visualization to traditional dialectology, this research enterprise was initiated by Séguy (1971) and has been taken up most actively in Salzburg (Bauer 2009; Goebl 2006; 1982) and Groningen (e.g. Heeringa & Nerbonne 2013; Nerbonne 2006; Heeringa 2004). At its core, quantitative dialectology collects information about a large number of individual linguistic features – often in the form of linguistic atlas data – to develop measurements of distance between individual dialects. These dialects are usually defined in terms of geography, although more recent work (Ghyselen & De Vogelaer 2018; Wieling et al. 2011) successfully incorporates information about sociolinguistic variation.

Similar methods have recently found productive application in English variation studies under the label of *corpus-based dialectometry* (Szmrecsanyi 2013). Research of this kind capitalizes on the availability of naturally occurring language data instead of relying on survey-based atlas information. Such data may come in the form of available corpora (Szmrecsanyi 2013) or may be collected by the individual researchers themselves (e.g. Grieve 2016). The most comprehensive efforts in this regard to date are Szmrecsanyi's (2013) study of spatial patterns in British English (BrE) morpho-syntactic variation and Grieve's (2016) analysis of American English (AmE) dialect regions on the basis of letters to the editor from 240 cities.

While these two examples demonstrate the advanced state of dialectometric research on the two globally dominant varieties BrE and AmE, less attention has been paid to global patterns of variation. In a series of articles drawing on information from the World Atlas of Variation in English (Kortmann & Lunkenheimer 2013a), Kortmann and Szmrecsanyi

(2011) establish different "typological profiles" for the global varieties of English. Their findings suggest that contact history is the most powerful distinguishing parameter, with geography only playing a secondary role (Szmrecsanyi & Kortmann 2009b: 1658). However, these studies rely heavily on categorical information in the form of expert testimony. A full centering of the lectometric enterprise in World Englishes on usage-based data is yet to be achieved (although see Heller et al. 2017; Szmrecsanyi et al. 2016 for developments in this direction). Given the existence of carefully compiled, representative corpora for and the role of geography in World Englishes research, it is highly appropriate at this point to develop such a corpus-based, aggregate perspective.

1.1.3 MD Analysis

In comparison to the studies mentioned above, however, one further aspect is worth emphasis: the role of register in structuring linguistic variation. One of the chief merits of the ICE project is its meticulous sampling framework, which represents a large and carefully defined range of situational-functional registers. Until now, dialectometry has largely taken a pragmatically agnostic position towards register variation. The typical practice is to restrict the analysis to one specific text type, such as letters to the editor (Grieve 2016) or interview data (Szmrecsanyi 2013). This practice implies a negligible role of register that is at odds with findings from ICE-based analyses. Summarizing one case in point, the development of modals and semi-modals in BrE and AmE, Mair (2015b: 139) concludes that "genre remains as the most important source of statistical noise."[2]

Rather than treating register as a potential threat contaminating the analysis, the present study proposes to investigate it in equal terms alongside geographic variation. The appropriate theoretical apparatus for this perspective is provided by the research tradition of MD analysis developed by Biber (1988). In terms of method, MD analysis shares much with dialectometry in that both capitalize on patterns of co-variation among a large amount of linguistic features in order to develop a bird's eye view of the total variation found in the data. The difference is that the MD approach places emphasis on the situational-functional context in which a given instance of language takes place, viewing linguistic differences

[2] For a discussion of register and related terms such as "genre" and "text type," see Section 2.2.

primarily as functions of different situations. For instance, the co-presence of interlocutors in a face-to-face conversation explains the frequency of linguistic features such as personal pronouns or clause-based structuring of discourse compared to these features' relatively low frequency in technical writing. Similarly, to take another example, the dense and abstract packaging of information in academic prose compared to narrative fiction is reflected in higher rates of occurrence of nominalizations and passive-voice constructions.

MD research is explicit in its assumption that situational-functional properties like the above are the fundamental explanatory factors of linguistic variation. Biber and Finegan (1994a: 10) go as far as claiming that "patterns of dialect variation are derivable from ... more basic patterns of register variation" and Biber (2014; 1995) shows that certain register differences apply across genetically and typologically unrelated languages. Considering dialectometry and MD analysis together, then, presents a picture of fundamental theoretical differences despite methodological commonalities. The present-day corpus landscape allows for an attempt at unifying the theoretical differences within an analysis that investigates geographic and register variation at the same time, without giving one a priori precedence over the other.

1.2 Research Objectives

Broadly speaking, then, the present study addresses the following research questions:

- How can the observable patterns of variation in a register-stratified corpus of World English be interpreted in terms of underlying dimensions they represent?
- To what extent are different national varieties of English characterized by linguistic unity and/or divergence, and how effective are the theoretical models in the World Englishes literature at accounting for such patterns?
- What is the respective explanatory contribution of register and variety in accounting for the variation observed in the corpus data?

With these general research goals outlined, some qualifying statements are in order. First, there is an additional extralinguistic dimension that correlates with linguistic variation in highly regular ways but which cannot be addressed systematically in the present study, namely: social meaning. Whether following a basic unidimensional logic of prestige or a more

multifaceted, constructionist theory of social meaning, the sociolinguistic record gives overwhelming evidence of the systematicity with which linguistic and social variables are related. The characteristics of the corpora used in the present book, however, do not allow for an easy incorporation of demographic data into the analysis. Therefore, very little can be said about social meaning in this study. It is my hope, however, that the method and some key findings developed here will be able to contribute to sociolinguistic studies in the future.

Second, the analytical perspective adopted here is concerned with the linguistic forest rather than individual trees. The broad-brushstroke development of fundamental dimensions of variation sacrifices detailed, contextualized attention to individual linguistic tokens. The vantage point proposed in this book, however, is by no means meant to replace detailed documentation and description of individual features, varieties, or speech communities. The level of ethnographic and structural insight gained by the fieldworker engaging one group of speakers cannot be achieved with the abstract, statistical methods presented here. Rather, the present study builds on such insights and hopes to establish a framework that helps researchers of individual varieties contextualize their findings within a wider picture of structured variation in English worldwide.

1.3 Outline

The remainder of the books is structured as follows:

Chapter 2 ("The World of English: Variation in Geography and Register") provides an introduction to the object of analysis – structural variation across a range of communicative situations in English worldwide. Section 2.1 presents the global spread of English in the early twenty-first century and discusses the available frameworks for understanding relations within the English language complex (McArthur 1998). The case is made that there is a disconnect between the rich level of theorizing and a lack of comprehensive empirical-linguistic accounts of differentiation in World Englishes research. Section 2.2 addresses the range of register variation documented in the English language. A brief discussion and definition of terms is followed by a sketch of the development in this area of research from a broad binary distinction between spoken and written language to the situationally, functionally, and structurally nuanced understanding available today. Section 2.3 outlines how register and geography interact in structuring linguistic variation and how, particularly in the study of

World Englishes, one may act as a confounding factor in the analysis of the other. Finally, Section 2.4 provides the rationale for including Twitter discourse as part of the empirical data in the study.

Chapter 3 ("Quantifying Linguistic Variation") introduces perspectives on the analysis and quantification of linguistic variation, comparing variationist sociolinguistics (Section 3.1), corpus-based text linguistics (Section 3.2), and the multifeature, aggregational approach associated with dialectometry and MD analysis (Section 3.3). The choice of the MD framework is justified and the central steps involved in such an analysis are outlined.

Chapter 4 ("The Space of Variation in the Present Study") discusses the data and method applied in the present study. Section 4.1 gives a concise overview of the ICE corpus project and the ten national sub-corpora that enter the analysis as well as describing how the Twitter (TwICE) corpus was sampled. Section 4.2 introduces the catalog of 236 linguistic features extracted from the corpus data. The chapter concludes with Section 4.3, in which the individual steps and parameter settings of the statistical procedure are described and a bird's eye view of the resulting space of variation is given.

Chapter 5 ("General Situational Dimensions of Variation") presents the first three dimensions of linguistic variation developed in the statistical analysis. The interpretation of these dimensions can largely be derived from differences in modality and attendant communicative-situational properties.

Chapter 6 ("Register-Specific Dimensions") presents the next three dimensions. These differ from those presented in Chapter 5 in that they are dominated by a single register, i.e. they show particularly high values for one specific kind of corpus text.

Chapter 7 ("Dimensions with Other Patterns of Distribution") discusses the remaining four dimensions, which are less easily subsumed under a logic of either modality or conventions of a narrowly defined register. For each of these, explanations are provided that have to do with drifts in discourse conventions, cultural differences, and grammatical peculiarities across varieties of English.

Chapter 8 ("Discussion: Feature Space and Geographical Space") returns to the empirical, theoretical, and methodological questions raised in Chapters 2 and 3. Section 8.1 summarizes the descriptive picture that emerges from the ten dimensions developed in the analysis and puts the space of variation in the present study into systematic theoretical relation to the extant literature. Section 8.2 addresses the influence of variety and

register in accounting for linguistic variation. It provides an assessment of the descriptive utility provided by different World Englishes models, followed by a critical perspective on the relative predictive weakness of variety compared to register when accounting for the total amount of variation in the data.

Chapter 9 ("Conclusion") retraces the main analytical steps in the book and the theoretical insights gained from the analysis and provides an outlook for future research conducted in a similar spirit.