17 Corpus-Based Approaches to Dialect Study

BENEDIKT SZMRECSANYI[1] AND LIESELOTTE ANDERWALD[2]

[1] KU Leuven
[2] University of Kiel

17.1 Introduction

Corpus linguistics is a methodology that draws on (more or less) systematic collections of naturalistic, machine-readable texts to make claims about linguistic phenomena and/or linguistic variation (for textbooks see, e.g., Biber 1998; Friginal and Hardy 2014; McEnery, Xiao, and Tono 2006). Thus unlike other methodologies in linguistics—for example, those that rely on experimental data, or on elicited linguistic knowledge, or on intuitions (either the linguist's own or somebody else's)—corpus linguistics is the methodological outgrowth of the usage-based turn in linguistics. This is because what is of interest in corpus linguistics is what language users *do* with language (that is, their *behaviour*), not what they *know* (or *think* they know) about language. There are many different kinds of corpora: they may contain written or spoken (transcribed) language, modern or historical texts (or both), standard or non-standard language, adult language or child language, native language or learner language, and so on. It is clear that the sort of material included in a corpus will constrain the range of research questions that can be asked on the basis of the material. Here is a little taster of the sort of issues that can be tackled in a corpus-based approach: let X be some linguistic feature (a morphological marker, a lexical item, a grammatical construction, and so on) in which the researcher is interested. Corpus study may then address the following questions, among others: How—in which contexts, in which way, subject to which restrictions—is X used in a given corpus? Comparing two or more corpora sampling different registers (e.g., conversation versus academic prose), in which register is X more frequent? Comparing two or more corpora sampling historical stages (e.g. nineteenth-century English versus twentieth-century English), has X become more or less frequent over time? Comparing two or more corpora sampling different geographic language varieties, in which variety is X more frequent, that is, more widely used? Using a sociologically annotated corpus, is X more frequently used by male or by female speakers, by younger or by older speakers, by speakers from lower or from higher social classes?

A little case study may illustrate some of these points. Most native speakers of English will have intuitions about the negator *ain't*. What can corpora tell us about the contexts in which *ain't* occurs? A query on the *Corpus of Contemporary American English* (COCA)

The Handbook of Dialectology, First Edition. Edited by Charles Boberg, John Nerbonne, and Dominic Watt.

(available, like all web-based corpora mentioned in this section, on Mark Davies' corpus portal at http://corpus.byu.edu; see Davies 2010) reveals that *ain't* occurs as negated form of *have*, as in (1); as negated form of *be*, as in (2); and as negated form of *do*, as in (3).

1. Age *ain't* got nothing to do with living or dying (COCA Bk:EdgeDarkWater)
2. If it *ain't* broke, why fix it? (COCA NPR_TalkNat)
3. Rose, you *ain't* see she's a woman? (COCA Bk:GirlGolden)

Also, COCA samples a number of different registers—spoken registers, fiction, magazine prose, newspaper prose, and academic prose. It turns out that in the COCA material, *ain't* is most popular in fiction (99 occurrences per million words [pmw]); *ain't* is least popular in academic prose, where it occurs only 3 times pmw. COCA is moreover a so-called monitor corpus—the earliest material it contains dates from the 1990s, and the corpus is being continually updated. A look at the diachronic frequency trajectory of *ain't* reveals that the form is on the decline in COCA: from 39 occurrences pmw in the 1990–1994 period to 22 occurrences pmw in the 2010–2012 period. What about regional variation? According to the *Corpus of Global Web-Based English* (GloWbe), *ain't* is most widely used in U.S. American English (frequency: 25 occurrences pmw), and least widely used in Pakistani English (frequency: < 3 occurrences pmw). British web-based texts take the middle road: here *ain't* occurs about 13 times pmw. Thus, contrary to what some Britons may believe, *ain't* is not an Americanism. Yes, the marker is very popular in American English, but we also do find it in British varieties of English. In fact, we know from corpus analysis that *ain't* is used all over England (Anderwald 2002, 149), and especially in traditional dialects in the South of England (Szmrecsanyi 2013, 56–58). In the realm of corpus-based *dialect* study, relevant corpora typically consist of orthographically transcribed interviews with dialect speakers, similar to sociolinguistic interviews that are customary in variationist sociolinguistics. In the remainder of this contribution, our take on dialect study is restricted to the study of TRADITIONAL DIALECTS. We essentially follow Trudgill (1990, 5) in defining traditional dialects as follows: "Traditional dialects are what most people think of when they hear the term dialect, spoken by (in Western societies at least) fewer and fewer people in 'remote and peripheral rural areas'." This is another way of saying that we exclude from consideration corpus-based work on variation between standard varieties (e.g., British English versus American English, Netherlandic Dutch versus Belgian Dutch), global varieties (of English, Spanish, etc.), and we will also not be dealing with "urban" or "social" dialectology, which is primarily concerned with sociolinguistically conditioned variation.

Of course, corpus-based dialect study shares many methods with neighboring disciplines. SYNCHRONIC (VARIATIONIST) SOCIOLINGUISTICS, also sometimes referred to as *social* or *urban dialectology*, is methodologically essentially the same as quantitative dialectology (the difference lying in the criteria chosen for sampling). Substantially, the focus is not typically on geographic variation, even though geography is occasionally considered (e.g., Tagliamonte, Smith, and Lawrence 2005). QUANTITATIVE TEXT LINGUISTICS in the spirit of, e.g., Biber (1988) or Mair (2006), is primarily concerned with text frequencies of linguistic phenomena in usage data. As such, quantitative text linguistics overlaps methodologically with corpus-based dialect study, especially when it comes to the rigorous methodology that guides, or should guide, corpus compilation. However, quantitative text linguists have not traditionally taken an interest in dialect data. HISTORICAL CORPUS LINGUISTICS also has methods and substance in common with corpus-based dialectology. Finally, because corpus-based dialectology relies on (transcribed) interactive interviews, it is also informed by methods and interpretational frameworks developed in DISCOURSE AND CONVERSATION ANALYSIS. For example, factors like repetition and persistence are also relevant in the

analysis of dialect corpora (Szmrecsanyi 2006). Corpus-based dialect study shares with other corpus-based approaches a focus on morphology, grammar, and discourse pragmatics; corpus-based studies in phonetics and phonology are, by contrast, considerably less widespread (but see, e.g., Rácz 2012). The reason for this bias is that many corpora currently available contain written material. And even those corpora that sample spoken language more often than not digitize orthographically transcribed words, which is convenient as long as one is not interested in pronunciation. We will come back to this problem in Section 17.6, "Future Directions."

A recent example of a corpus-based dialect study is Anderwald's (2009) study of non-standard past tense forms (e.g., past tense *give, come, sung, drunk*, or *catched*) in traditional British English dialects, based on the *Freiburg Corpus of English Dialects* (FRED) (presented in more detail in Section 17.2). Because of the limits imposed by a finite corpus, attention had to be restricted to variable verbs that occur relatively frequently in the corpus material. Even so, however, a fine regional differentiation would have resulted in many empty cells for individual locales (or for individual lexemes). Rather than aggregate data across lexemes, Anderwald chose larger regional subdivisions. This is an example of the typical trade-off situations in comparative work, where breadth of coverage and depth of investigation (be it geographical, historical, linguistic, or other) cannot be simultaneously achieved.

Even given this trade-off, in doing comparative dialect studies we already make a number of important assumptions, most importantly that of equivalence. We assume that we are investigating phenomena that can in fact be compared across dialects (in Anderwald's case: some dialects may not have a morphological category of PAST TENSE). Under such circumstances, the research question should probably be changed into an onomasiological one (e.g., "how is reference to past events expressed in dialect X?"), and researchers will have to define clearly what they regard as "alternative ways of saying the same thing."

Comparative work across dialects also assumes that the data we work with are formally equivalent, for example, are transcribed consistently across locales, are equivalent in size and "vernacularity," be of acceptable audio quality, and that speakers are comparable in terms of parameters such as social class, gender, age, education, or ethnicity. In fact, Wolk (2014) demonstrates that sociolinguistic imbalances like this can in fact distort results in a corpus-based dialectology approach—even in dialect corpora whose design is overall quite (but not perfectly) homogeneous and balanced.

In summary, a number of assumptions, prerequisites, and difficulties characterize corpus-based approaches to dialect study:

- The target phenomena have to be textually relatively frequent, as the absence of features in corpus material is hard to interpret.
- The data subject to analysis must be relatively homogeneous across locales (in terms of sociolinguistic parameters, but also in terms of length, quality of recording, quality of transcription, etc.).
- Transcription must be as reliable, faithful, and internally as consistent as possible. At the same time, a certain degree of regularization is indispensable, to enable the identification of non-standard forms in the material.
- When corpus-based dialectologists adopt the variationist method, it is imperative to ensure that the variants subject to study are really equivalent ways of saying the same thing (see, e.g., Cheshire 2005; Lavandera 1978).
- Forms that look similar do not necessarily have to be the same (sometimes called camouflage constructions, e.g., Spears 1982).

Some of these issues will be revisited in Section 17.5 below.

17.2 Examples of Dialect Corpora

The research community is not exactly drowning in publicly available corpora that sample traditional dialects, as per our definition in Section 17.1, and such corpora as exist mostly cover dialects of English. For example, the *Freiburg Corpus of English Dialects* (FRED) (see Hernández 2006 for a manual) is a (more or less) synchronic (but historical) corpus—most of the material was collected in the 1970s—which covers several dialect areas in Britain, concentrating on traditional dialect speakers, many of them older males (i.e., it is not designed to be representative in terms of social class, gender, or age). FRED has been orthographically transcribed. The following is a representative extract (text CON_005, county Cornwall, Southwest of England):

Interviewer: ...recording of Wallace Jeff Baggerly of Porthmeor Farm, near Zennor, was made on the fifth of September nineteen seventy-eight. When were you born?
Informant: in nineteen hundred and four. Seventeenth of December.
Interviewer: And had your -- and your family had lived here…?
Informant: Yes, my father was born here. Not in this house!
Interviewer: No.
Informant: But in the old house, you know. And so was my grandfather.
Interviewer: Yeah. And that's the old house across the road, is it?
Informant: No, no, gone. Used to be here. You know, under this, see – or, under – somewhere, like. And that 's gone. And let 's see. His father again come up from down Lower Porthmeor.
Interviewer: Mhm.
Informant: Now I don't know quite how long they 'd been here, but they come from St. Hilary, somewhere. To start with, if you understand what I mean. That might 've been my great-grandfather's (pause) grandfather, perhaps. Somebody come, and a – with a baby. And that was one of the oldest old men that was here 'round, you know, but I couldn't tell you exactly which generation, you know.
Interviewer: No.
Informant: I do know my great-grandfather was born down Lower Porthmeor, and he had uh, one, two, three brothers. I knew them by name but I didn't know them (unclear) that well Uncle Richard and Uncle Jack, and Uncle Albert, you know, they was uncles to my grandfather, see, and, and you pick up the (unclear) sayin' from years ago.
Interviewer: Yeah.
Informant: That old house there, I did hear a great-uncle of mine say that he could mind somebody living in en. And that, he was, he was little. Well this is Tim's house to this day.

Note, for example, how the informant uses non-standard *was* in plural contexts (… *they was uncles* …). Based on this and the other texts in FRED, Szmrecsanyi (2010) explores the regional distribution of non-standard *was* across the major dialect areas in Great Britain, and finds that the feature does not have a significant geographic distribution overall, although it does tend to be more frequent in traditional dialects in England than in traditional dialects in Scotland (see the box plot in Figure 17.1).

The full version of FRED is available to researchers and visiting scholars at the University of Freiburg. The 1-million word sampler version of FRED (FRED-S; see Szmrecsanyi and Hernández 2007 for a manual) is publicly available and comes with part-of-speech annotation.

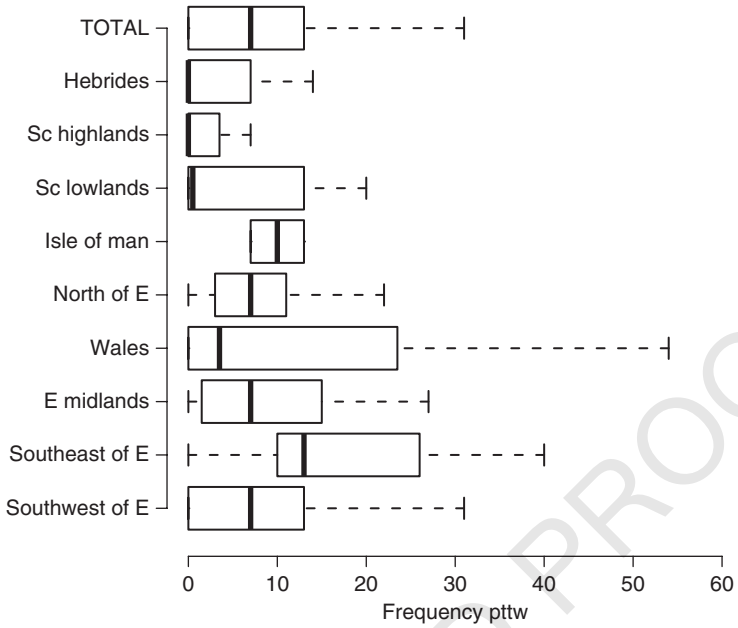

Figure 17.1 Box plot depicting frequency variance of non-standard *was* by dialect region in FRED (Szmrecsanyi 2010, 48).

In the *Diachronic Electronic Corpus of Tyneside English* (DECTE; see Corrigan, Mearns, and Moisl 2014), the locality is—by contrast to FRED—held constant: all material comes from Tyneside, but the time axis is extended, as is the social coverage. DECTE contains material from the 1960s, the 1990s (Milroy, Milroy, Hartley, and Walshaw 1994), and from the first decade of the twenty-first century. This diachronic structure allows researchers to compare dialect forms over time, and to a degree to comment on their social distribution (cf. Beal, Burbano-Elizondo, and Llamas 2012). Features that have been analyzed so far on the basis of DECTE include negation (Beal and Corrigan 2005), relativisation (Beal and Corrigan 2002; 2006), intensifiers (Barnfield and Buchstaller 2010), and phonetic variation (Corrigan, Mearns, and Moisl 2014).

A corpus that is sometimes used to investigate the regional distribution of widespread features of spoken English is the spoken material in the British National Corpus (BNC, cf. Aston and Burnard 1998), with about 5 million words of unmonitored everyday speech. However, the material was not originally intended to be regionally representative, so that both size, quality, and social make-up of the material differ significantly across regions, and the transcription was not produced by linguists. Nevertheless, studies based on the BNC have produced interesting regional results, for example, for features such as pseudopassives (Klemola 1999), the system of the English verb phrase quite generally (Sampson 2002), various features of negation (Anderwald 2002), or ditransitives (Gerwin 2013).

Important contrastive work is also carried out by sociolinguists and dialectologists on materials that are typically not made available to outsiders. Tagliamonte has collected (or supervised the collection of) corpora of York English (e.g., Tagliamonte 1998; 2001), Devon English (e.g. Godfrey and Tagliamonte 1999), Toronto English (e.g., Tagliamonte and D'Arcy 2007), or Samaná English (e.g. Poplack and Tagliamonte 2001), and typically feature analyses are compared across several of these materials (cf. Tagliamonte 2012b). Tagliamonte has also

collected materials from isolated, peripheral dialect communities in Northern Britain and Ireland in her *Roots of English* project (Tagliamonte 2012a), specifically from localities in the Northwest of England, Lowland Scotland, and Northern Ireland. Her morphosyntactic analyses include subject-verb concord, the use of adverbs without *–ly*, negative versus auxiliary attraction, relative pronouns, *that*-complementation, *for to* infinitives, future reference, the competition of past tense and present perfect, modals of obligation, possessive *have* versus *have got*, discourse marker *like* and general extenders (e.g., *and stuff like that*) (all in Tagliamonte 2012a). Rather than comparing surface structures or relative frequencies, in her "comparative sociolinguistics" approach (Tagliamonte 2012b) multivariate analyses serve to uncover the relevance of extralinguistic and intralinguistic constraints, and it is the order and strength of these constraints—rather than absolute or relative frequencies—that serve as the point of comparison across dialects (already in Poplack and Tagliamonte 2001). (Note also that many of the features she investigates are actually (variable) features of Standard English, rather than dialectal in the strict sense.)

As for dialects of languages other than English, corpus resources are still comparatively rare, but the situation is improving. Consider, for example, the *Corpus Gesproken Nederlands* (CGN; see http://lands.let.ru.nl/cgn/ehome.htm), which can be utilized for dialectological analysis (e.g., Hoste, Gillis, and Daclemans 2000); or the *Estonian Dialect Corpus* (see http://www.murre.ut.ee/murdekorpus/and Uiboaed *et al*. 2013); or the "Phonologie du Français Contemporain" [Phonology of contemporary French] project (see http://www.projet-pfc.net/and Durand 2006); there is also a project on "Phonologischer Wandel am Beispiel der alemannischen Dialekte Südwestdeutschlands im 20.Jahrhundert" [Phonological change in Alemannic dialects in Southwest Germany], which among other things draws on spontaneous-conversational corpus material to explore dialectal change (see Auer, Baumann, and Schwarz 2011; Streck and Auer 2012).

17.3 Research Questions

In our overview above several research questions have already been implicit. Making them explicit, we distinguish in what follows (1) geolinguistics proper, serving synchronic, diachronic or typological interests, (2) aggregate geolinguistics, (3) the discovery of constraint rankings, and (4) data mining.

The perhaps most time-honored motivation for performing comparative dialectology is the quest for areal patterns, something we may call GEOLINGUISTICS PROPER. Just as we are able to draw isoglosses of phonetic variants on maps, we can proceed in the same way for features of morphosyntax, and thus directly correlate linguistic and geographical information. The underlying motivation for discovering areal patterns might be manifold. As Kehrein (2012) has pointed out, areal patterning can be taken as indicative of different paths of language change from a (purported) original single source; geolinguistic patterns are then of interest in a structuralist, intralinguistic way, indicating the (synchronic) range of variation possible in one language. Kehrein quotes German Neogrammarian linguists of the nineteenth century as examples of this research paradigm. A more recent, but essentially very similar avenue of interpreting areal patterns is found in interpretations of a functional-typological kind (as in Kortmann 2002; Kortmann, Pietsch, Herrmann, and Wagner 2005). Here the interest is in discovering the breadth of variation synchronically, and in identifying dominant and minority patterns. Once dominant patterns are identified, they are then interpreted in terms of their (systemic or psycholinguistic) function, and often compared with patterns observable in other languages. In a cross-dialectal and cross-linguistic comparison, for example, it becomes quickly apparent that many features (such as multiple negation, the lack of marking adverbs, or the use of invariant tag questions) of non-standard English are

perfectly "normal" in a world-wide perspective, and that it is Standard English which stands out in not following these unmarked typological trends.

In a related manner of interpretation, geolinguistic patterns always held interest for historical linguistics, because they can also be read as indicating individual stages in diachrony. Geolinguistic variation is then seen as making visible the process of language change, as relic areas preserve older forms. Many maps in the Survey of English Dialects (SED) have been interpreted in this way (cf. the introductory chapter in Orton, Sanderson, and Widdowson 1978). The interest might then shift to extralinguistic factors that may have caused the observed distributions, moving into socio-cultural terrain. Thus, researchers have asked about settlement patterns, as in the Linguistic Atlas Projects of the United States (e.g., Kurath 1949), where fine-grained dialect divisions on the Eastern seaboard "fan out" toward the west into a more homogeneous area. Transport and communication networks have been invoked in the interpretation of relic areas (for the East Anglian Fens, cf. Britain 2002), or in the geographical spread of new variants (cf. Labov 2010, chapter 10). In England, the status of London and the wider Southeast as a source of leveling, and a source of innovations that spread to the rest of the country, has also been apparent through consistent patterns in dialect maps (e.g., for historical phenomena like H-dropping, the loss of postvocalic R, or L-vocalization, all in Orton *et al.* 1978). In reverse, geographical distance, but also a sociocultural sense of isolation have been invoked to explain the status of relic areas such as the Northeast of England (cf. Beal 2004). However, even in mainstream dialectology more modern conceptualizations of geography as imagined spaces (e.g., Zelinsky 1973) have not really entered mainstream publications yet.

The research cited in the foregoing discussion is primarily concerned with the geolinguistic patterning of individual phenomena. However, dialect areas are typically constituted by the bundling of isoglosses. More sophisticated methods to conduct aggregate corpus-based dialectology are now available to model variable rather than categorical features, and several rather than single ones (see, e.g., Szmrecsanyi 2013; Wolk 2014). Such methods may address new research questions, such as: To what degree does dialectal similarity or dialectal distance correlate with geographic distance? What sort of geographic distance counts—as-the-crow-flies distance, travel time, travel distance, and so on? Which features contribute to specific geographic patterns, or, in Szmrecsanyi's words, "how do features gang up to create layered areal patterns" (Szmrecsanyi 2013, 137)?

The comparative sociolinguistics paradigm (Tagliamonte 2012a,b), which empirically relies exclusively on usage (corpus) data, has already been mentioned. Here, instead of the identification of areal patterns, dialects are compared quantitatively for their underlying CONSTRAINT RANKINGS that determine the observed surface patterns. The quantitative analysis thus yields qualitative differences, and identical rankings are taken to indicate close historical relations, typically supplemented by socio-historical evidence (for a critique, cf. Pietsch 2012).

The idea of identifying "underlying" patterns is perhaps reminiscent of the generative enterprise. However, it has to be said that a systematic investigation of dialect differences has not been at the core of Generative Grammar. While different language systems are occasionally included in generative arguments (e.g. Halle and Mohanan 1985), this is typically not based on preceding quantitative analyses, but on introspection, or access to individual informants. This is understandable, given the deep-seated scepticism in generative circles concerning the validity of corpus linguistics (Chomsky 1956; 1957; Miller and Chomsky 1963). Where dialect material is collected and compared, the interest is essentially in uncovering differences in the deep structure (e.g., Adger and Smith 2005 for a northeastern Scottish community; Henry 1995 for Belfast English; Tubau Muntañá 2008 for multiple negation in Britain), and qualitative differences are usually taken as more important than quantitative ones.

17.4 Methods

Against the backdrop of the research questions discussed in the previous section, three major methods in corpus-based dialect study may be distinguished: (1) qualitative example mining, (2) quantitative single-feature studies, and (3) quantitative multi-feature studies. Let us discuss these in turn.

Qualitative example miners tap into dialect corpora to obtain evidence of the attestedness of particular linguistic features in particular dialects. Relevant examples include Henry (1995) for Belfast English constructions, or Tubau Muntañá (2008) for multiple negation in FRED.

quantitative single-feature study uses quantitative methods to investigate one feature at a time (see Nerbonne 2009, 176–1177 for a critical discussion); the contribution by Anderwald (2009) mentioned above, but also those studies collected in Kortmann, Herrmann, Pietsch, and Wagner (2005) or Hernández, Kolbe, and Schulz (2011) are representative of recent work in this spirit on the grammar of traditional British English dialects. We can more specifically distinguish two variants of this approach: frequency-focused single-feature study, and constraint-focused single-feature study. In frequency-focused single-feature study, dialectologists determine usage frequencies of particular features. In this endeavour, increased frequency is typically considered a proxy of a feature's entrenchment and/or overall importance in a particular dialect grammar. Representative examples of this approach include Anderwald (2009) discussed in Section 17.1, or Herrmann (2005) on relativisation strategies in FRED. Constraint-focused single-feature study is the sort of multivariate approach that characterizes variationist sociolinguistic work by Tagliamonte and collaborators mentioned above. In this line of analysis the question is "When dialect speakers have a choice between two ways of saying the same thing, which factors (language-internal or language-external) constrain their choice"? Addressing this question necessitates using multivariate analysis methods such as binary logistic regression (e.g., Varbrul). Representative work in this tradition includes Pietsch (2005), who is interested in the conditioning of verbal agreement patterns in Northern dialects of English, or Tagliamonte and Smith (2005), who study complementizer *that* retention and omission in British English dialects. Even though constraint-focused variationist (socio)linguists do not necessarily consider themselves corpus linguists, since their work is based on collections of authentic language usage (i.e., corpora), we do consider their work relevant here.

In quantitative multi-feature study (a.k.a. corpus-based dialectometry), analysts base claims not on the distribution of one particular feature, but of many. Quantitative multi-feature study thus adopts dialectometrical methods (see Chapter 7). The goal is to obtain a more robust geolinguistic signal; this signal can then be projected to geography in sophisticated exploratory maps, and/or correlated with language-external measures such as geographic distance. Unlike in traditional dialectometry (Séguy 1971; Goebl 1982; Nerbonne, Heeringa, and Kleiweg 1999), the primary data are not contained in dialect atlases or surveys but come from dialect corpora. Two ways of doing corpus-based dialectometry may be distinguished: top-down corpus-based dialectometry, and bottom-up corpus-based dialectometry. The top-down approach first defines a feature catalogue, then establishes frequencies (Szmrecsanyi 2013) of or probabilities (Wolk 2014) associated with these features, and subsequently calculates a joint measure of pairwise linguistic distances between the dialects considered. For example, Szmrecsanyi (2013) explores the extent to which grammatical variation in British English dialects is structured geographically—and thus, is sensitive to the likelihood of social contact. The study combines corpus-based variation studies with aggregative-dialectometrical analysis and visualization methods. This synthesis is desirable for two reasons. First, dialects are multidimensional, and hence call for aggregate analysis techniques. Second, compared to linguistic atlas material, corpora yield a

more radically usage-based frequency signal. Against this backdrop, Szmrecsanyi calculates an aggregate measure of dialect distance based on the discourse frequency of 57 morphosyntactic features, such as multiple negation, non-standard verbal *-s* (e.g., *so I says, What have you to do?*), or non-standard weak past tense and past participle forms (e.g., *they knowed all about these things*) in FRED (see Section 17.2). The ultimate aim is to reveal large-scale patterns of grammatical variability in traditional British English dialects. Referring back to the research questions in Section 17.3, Szmrecsanyi's study shows that it is impossible to find in England a clearly demarcated Midlands dialect area on grammatical grounds, and that travel time is a better predictor of linguistic distance than as-the-crow-flies geographic distance. In a broadly similar vein, Grieve (2009; see also 2011, 2012)[1] is interested in regional grammatical variation in American English. He defines a feature catalogue spanning 45 (standard English) grammatical variables, and examines their usage rates in a huge corpus of letters to the editor in 200 cities from across the United States. Contrary to what old-school dialectologists may have suspected, Grieve demonstrates that his rather unorthodox, written material indeed exhibits geolinguistic patterns.

In BOTTOM-UP CORPUS-BASED DIALECTOMETRY, by contrast, features are not defined *a priori*, but are allowed to emerge in a data-driven fashion. In this spirit, Wolk (2014) uses a part-of-speech-annotated version of FRED, and develops a probabilistically enhanced method (based on Nerbonne and Wiersma 2006) that draws on part-of-speech bigram frequencies (e.g., sequences of determiner-noun) to calculate an aggregate measure of dialect distance. The resulting geolinguistic signal is weaker than that yielded by top-down approaches, but it does uncover dialectologically meaningful areal patterns.

17.5 Issues and Problems

We have already implicitly hinted at potential problems and issues with the corpus-based study of non-standard materials, especially those to do with transcription and normalization regimes. Consistency of transcription, devising normalized spellings for non-standard items that enable some computer-readability without making too many theoretical assumptions, and the internal make-up of subsamples—in particular interaction between social and regional variation—are potential problem areas that researchers have to be aware of. We have also pointed out that all corpora (through being finite resources) impose a qualitative limit on what can be investigated: only features that are frequent enough can be included in comparative analyses. Arbitrary cut-off points can here lead to inter-researcher differences. What was investigated by Anderwald (2009) (past tense *drunk, sung, rung*) was excluded by Szmrecsanyi (2013) on the grounds that this feature just did not make the lower frequency threshold. The role of frequency also leads to the more general problem of how to deal with extremely rare, or even absent, features. The absence of a feature from a corpus can be a sign of its overall rarity: thus past or modal perfect progressive passive forms (e.g., *would have been being charged*) even in a huge Standard English corpus like COCA (see Section 17.1) are only attested four times, and we would thus expect that in smaller corpora these complex verb phrases, though no doubt in principle possible in English, would not occur at all. Although this problem related to size has increasingly been counteracted by building larger and larger corpora, the automation processes typically employed are not really feasible for non-standard materials, and corpora of several hundred million words like COCA are probably not a realistic goal when it comes to dialect material. This means that empirically, absence due to low text frequency is difficult to distinguish from truly ungrammatical forms (the problem of "negative evidence"). As dialectologically relevant examples, *was sat/stood* with progressive meaning, or resumptive relative pronouns (of the type *the house which he saw it*), despite being regularly cited in the dialectological literature, are not (or only very infrequently) found in FRED.

Other features are not only rare overall, but occur in such specific discourse contexts that the corpus material perhaps does not provide for their occurrence. Thus it has been observed that the English "hot-news" perfect or habitual constructions, which are both only used in pragmatically marked situations, are conspicuously absent from FRED (Anderwald and Wagner 2007). The reverse of this dependency on frequency may also be a problem, though, if a researcher only investigates what one can investigate safely, and thus lets the corpus material dictate the research question. Trivially counting for the sake of counting, possibly even without a working hypothesis, is an inherent danger in all corpus linguistics, and we only note it here for the sake of completeness, politely refraining from mentioning actual examples.

Particularly relevant for comparative dialectological work, we note the possible mistake of investigating as dialect (or even wider) universals what is trivially (namely historically) given in all varieties. Thus, in a review of comparative articles on a range of varieties of English around the world, Trudgill notes that "the fact that nonstandard dialects of English have many similarities … is really of no great interest … and to attempt to extrapolate universal principles out of the commonality is to credit the similarities with more importance than actually they have" (Trudgill 2013, 87), citing multiple negation, *there's* followed by plural NPs, or present participles in <-in> as examples.

Finally, we list here three more areas where we would claim that corpus-based dialect studies are probably not very useful, besides those rare and those pragmatically marked features noted above: the investigation of very local (i.e., areally skewed) features, the investigation of categorical features, and the investigation of features where surface similarities mask deeper differences (as noted above in Section 17.2). Areally skewed features, although perhaps in some respects the most interesting ones in terms of linguistic geography, do not lend themselves well to *comparative* quantitative analysis, mainly for technical reasons, because the resulting empty cells for many areas act as knock-out constraints in variable analyses. The same can be said for categorical (non-variable) features, which also do not lend themselves well to a comparative analysis of *variability*. As our final point, surface similarities that are due to underlying differences result in comparing apples and oranges, as we have noted above (although it is not always a trivial matter to find out what constitutes the apple, and what the orange …).

17.6 Future Directions

Corpus-based approaches to dialect study are currently being refined in the following ways. For one thing, the foregoing discussion mentioned social imbalances in corpus material as a nuisance factor in corpus-based dialect study. But as a matter of fact, corpus analysts are beginning to explore the exciting opportunities that corpora offer with regard to the interface between dialectology and sociolinguistics. In this line of work, geographically conditioned patterns still take centre stage, but thanks to corpora which do not exclusively sample non-mobile old rural males (NORMs) we can increasingly explore the extent to which geographic patterns are different when our attention is restricted to male or female speakers, old or young speakers, and so on. An exemplary study highlighting the potential of interface explorations along these lines is Heeringa and Hinskens (2014), who tap into a parallel corpus database and exploit social differences between their informants to study dialect change in the Dutch language area in apparent time.

Secondly, future work is likely to advance bottom-up approaches in the spirit of Nerbonne and Wiersma (2006) and Wolk (2014). Bottom-up corpus analysis is actually quite common in, for example, phraseology and collocation research, but in corpus-based dialectology the potential afforded by bottom-up analysis is as yet underexplored. The possibility of bottom-up analysis is actually what most radically sets apart corpus-based dialectology from dialectology

based on other data sources, such as dialect atlases (although some bottom-up approaches are used here, too, as in Kretzschmar's "self-organizing" maps, e.g., Kretzschmar 2011).

Third, observe that previous corpus-based dialectology research is overwhelmingly grammar-centred (much like corpus linguistics in general has a bias towards grammar and morphology, as we noted in Section 17.1). But many of the dialect corpora that this research draws on also provide audio material, which in most cases still awaits systematic phonetic analysis, both auditory and acoustic (along the lines of Grieve 2014). Lexis is likewise a neglected domain in corpus-based dialect studies, but this neglect is primarily due to the fact that lexical research requires large corpora, and conventional dialect corpora are simply not large enough for lexical analysis. (Also, of course, corpus-based dialectology was invented to address shortcomings in traditional dialectology, and traditional dialectology has always had a strong focus on lexis, besides phonetics.)

Fourth, corpus-based dialectology has so far only marginally interpreted its results in a wider societal context, and collaborations with cultural studies for informed sociocultural analyses are a desideratum. We can imagine fruitful combinations of corpus-based dialect study with

1. perceptual dialectology, which would allow a third kind of geographical distance to be included in the analysis, that is, the "perceived" distance between locales (as in Montgomery and Beal 2011);
2. linguistic anthropology, for example, for a study of attitudes, the covert and overt prestige of features, or for aspects of identity construction, especially at the group level (perhaps through an in-depth study of the meta-reflexive enregisterment of individual features in speakers' awareness of their own variety versus the varieties of others, along the lines of Johnstone, Andrus, and Danielson 2006 for Pittsburgh English);
3. social history, to address issues such as migration and settlement patterns, urbanization patterns, the spread of standard languages through education, and so on.

And finally, we note that there is an in principle well-known overlap between the sorts of research questions asked in dialectology, on the one hand, and in crosslinguistic typology on the other hand (see, e.g., the papers in Kortmann 2004). Recent years have seen some methodological convergence, in that some crosslinguistic typologists now increasingly rely on (parallel) corpus databases, instead of decontextualized reference grammars or individual expert informants (consider the papers in Szmrecsanyi and Wälchli 2014). It will be worthwhile to further explore these methodological interfaces, for the sake of developing a more unified discipline of geolinguistics, and to contribute to a unified study of intra- and crosslinguistic variation.

NOTES

1 Even though Grieve (2009) is not concerned with "traditional" dialects as defined in Section 17.1, we include this work in our survey due to its methodological innovativeness, and because his results nevertheless produce clear geographical patterns

REFERENCES

Adger, David and Smith, Jennifer. 2005. Variation and the Minimalist Program, In Cornips, L. & Corrigan, K. (eds.), *Syntax and Variation: Reconciling the Biological and the Social*, John Benjamins, Amsterdam & Philadelphia, pp. 149–178.

Anderwald, Lieselotte, 2002. *Negation in Non-Standard British English: Gaps, Regularizations and Asymmetries*, Routledge, London & New York.

Anderwald, Lieselotte. 2009. *The Morphology of English Dialects: Verb-Formation in Non-Standard English*, Cambridge University Press, Cambridge.

Anderwald, Lieselotte and Wagner, Susanne. 2007. FRED - The Freiburg English Dialect corpus, In Beal, J. C., Corrigan, K. P. & Moisl, H. (eds.), *Creating and Digitizing Language Corpora*, Macmillan, London, pp. 35–53.

Aston, Guy and Burnard, Lou. 1998. *The BNC Handbook: Exploring the British National Corpus with SARA*, Edinburgh University Press, Edinburgh.

Auer, Peter, Baumann, Peter and Schwarz, Christian. 2011. Vertical vs. horizontal change in the traditional dialects of southwest Germany. A quantitative approach. *Taal en Tongval* 63(1).

Barnfield, Kate & Buchstaller, Isabelle. 2010. Intensifiers on Tyneside: Longitudinal developments and new trends. *English World-Wide* 31: 252–287.

Beal, Joan C. 2004. "Geordie Nation": Language and regional identity in the Northeast of England. *Lore & Language* 17: 33–48.

Beal, Joan C., Burbano-Elizondo, Lourdes, and Llamas, Carmen. 2012. *Urban North-Eastern English: Tyneside to Teesside*, Edinburgh University Press, Edinburgh.

Beal, Joan C. and Corrigan, Karen P. 2002. Relativisation in Tyneside and Northumbrian English, In Poussa, P. (ed.), *Relativisation on the North Sea Littoral*, Lincom, Munich, pp. 125–134.

Beal, Joan C. and Corrigan, Karen P. 2005. "No, nay, never": Negation in Tyneside English, In Iyeiri, Y. (ed.), *Aspects of English Negation*, John Benjamins, Tokyo: Yushodo University Press and Amsterdam & Philadelphia, pp. 139–156.

Beal, Joan C. and Corrigan, Karen P. 2006. A tale of two dialects: Relativization in Newcastle and Sheffield, In Filppula, M., Klemola, J., Palander, M. & Penttilä, E. (eds.), *Dialects Across Borders: Selected Papers from the 11th International Conference on Methods in Dialectology (Methods XI), Joensuu, August 2002*, Cambridge University Press, Cambridge, pp. 211–229.

Biber, Douglas. 1988. *Variation Across Speech and Writing*, Cambridge University Press, Cambridge.

Biber, Douglas. 1998. *Corpus Linguistics: Investigating Language Structure and Use*, Cambridge University Press, Cambridge.

Britain, David. 2002. Diffusion, levelling, simplification and reallocation in past tense BE in the English Fens. *Journal of Sociolinguistics* 6: 16–43.

Cheshire, Jenny. 2005. Syntactic Variation and Beyond: Gender and Social Class Variation in the Use of Discourse-new Markers. *Journal of Sociolinguistics* 9(4): 479–508.

Chomsky, Noam. 1956. Three models for the description of language. *Transactions on Information Theory* 2: 113–124.

Chomsky, Noam. 1957. *Syntactic Structures*, Mouton, The Hague.

Corrigan, Karen P., Mearns, A.J. and Moisl, Hermann. 2014. Feature-based Versus Aggregate Analyses of the DECTE Corpus: Phonological and Morphological Variability in Tyneside English, In Szmrecsanyi, B. and Wälchli, B. (eds.), *Cross-Linguistic and Language-Internal Variation in Text and Speech*, Walter de Gruyter, Berlin, pp. 113–149.

Davies, Mark. 2010. More than a peephole: Using large and diverse online corpora. *International Journal of Corpus Linguistics* 15: 405–411.

Durand, Jacques. 2006. Mapping French Pronunciation: The PFC project, In Montreuil, J.-P. (ed.), *New Perspectives on Romance Linguistics*. Vol. 2: Phonetics, Phonology and Dialectology. John Benjamins, Amsterdam/ Philadelphia, pp. 65–82.

Friginal, Eric and Hardy, Jack. 2014. *Corpus-Based Sociolinguistics. A Guide for Students*, Routledge, New York.

Gerwin, Johanna. 2013. 'Give it me!': Pronominal ditransitives in English dialects. *English Language and Linguistics* 17: 445–463.

Godfrey, Elizabeth and Tagliamonte, Sali. 1999. Another piece for the verbal *-s* story: Evidence from Devon in southwest England. *Language Variation and Change* 11: 87–121.

Goebl, Hans. 1982. *Dialektometrie: Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie*, Österreichische Akademie der Wissenschaften, Wien.

Grieve, Jack. 2009. *A Corpus-Based Regional Dialect Survey of Grammatical Variation in Written Standard American English*, PhD Dissertation, Northern Arizona University.

Grieve, Jack. 2011. A regional analysis of contraction rate in written Standard American English, *International Journal of Corpus Linguistics* 16: 514–546.

Grieve, Jack. 2012. A statistical analysis of regional variation in adverb position in a corpus of written Standard American English.

Corpus Linguistics and Linguistic Theory 8: 39–72.
Grieve, Jack. 2014. A Comparison of Statistical Methods for the Aggregation of Regional Linguistic Variation, In Szmrecsanyi, B. & Wälchli, B. (eds.), *Cross-Linguistic and Language-Internal Variation in Text and Speech*, Walter de Gruyter, Berlin, pp. 53–88.
Halle, Morris and Mohanan, Karuvannur P. 1985. Segmental phonology of Modern English. *Linguistic Inquiry* 16: 57–116.
Heeringa, Wilbert and Hinskens, Frans. 2014. Convergence Between Dialect Varieties and Dialect Groups in the Dutch Language Area, In Szmrecsanyi, B. & Wälchli, B. (eds.), *Cross-Linguistic and Language-Internal Variation in Text and Speech*, Walter de Gruyter, Berlin, pp. 26–52.
Henry, Alison. 1995. *Belfast English and Standard English: Dialect Variation and Parameter Setting*, Oxford University Press, New York & Oxford.
Hernández, Nuria. 2006. *User's Guide to FRED*, University of Freiburg, Freiburg. URN:nbn:de:bsz:25-opus-24895, URL: http://www.freidok.uni-freiburg.de/volltexte/2489/.
Hernández, Nuria, Kolbe, Daniela and Schulz, Monika Edith. 2011. *A Comparative Grammar of British English Dialects: Modals, Pronouns and Complement Clauses*. Mouton de Gruyter, Berlin/New York.
Herrmann, Tanja. 2005. Relative Clauses in English Dialects of the British Isles, In Kortmann, B., Herrmann, T., Pietsch, L. and Wagner, S. (eds.), *A Comparative Grammar of British English Dialects: Agreement, Gender, Relative Clauses*, Mouton de Gruyter, Berlin/New York, pp. 21–124.
Hoste, Veronique, Gillis, Steven and Daclemans, Walter. 2000. A rule induction approach to modeling regional pronunciation variation, In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, Association for Computational Linguistics, pp. 327–333.
Johnstone, Barbara, Andrus, Jennifer and Danielson, Andrew E. 2006. Mobility, indexicality, and the enregisterment of 'Pittsburghese'. *Journal of English Linguistics* 34: 77–104.
Kehrein, Roland. 2012. Linguistic atlases: Empirical evidence for dialect change in the history of languages, In Hernández-Campoy, J. M. & Conde-Silvestre, J. C. (eds.), *The Handbook of Historical Sociolinguistics*, Wiley-Blackwell, Malden, MA & Oxford, pp. 480–500.
Klemola, Juhani. 1999. *Still sat in your car?* Pseudopassives with *sat* and *stood* and the history of non-standard varieties of English English. *Sociolinguistica* 13: 129–140.
Kortmann, Bernd. 2002. New prospects for the study of English dialect syntax: impetus from syntactic theory and language typology, In Barbiers, S., Cornips, L. and Kleij, S. v. d. (eds.), *Syntactic Microvariation*, Meertens Institute, Amsterdam, pp. 185–213.
Kortmann, Bernd (ed.). 2004. *Dialectology Meets Typology: Dialect Grammar from a Cross-Linguistic Perspective*, Mouton de Gruyter, Berlin & New York.
Kortmann, Bernd, Pietsch, Lukas, Herrmann, Tanja and Wagner, Susanne. 2005. *A Comparative Grammar of English Dialects: Agreement, Gender, Relative Clauses*, Mouton de Gruyter, Berlin & New York.
Kretzschmar, William A., Jr. 2011.The beholder's eye: Using self-organizing maps to understand American dialects, In Adams, M. and Curzan, A. (eds.), *Contours of English and English Language Studies*, University of Michigan Press, Ann Arbor, Mi., pp. 53–70.
Kurath, Hans. 1949. *A Word Geography of the Eastern United States*, University of Michigan Press, Ann Arbor, Mi.
Labov, William. 2010. *Principles of Linguistic Change*, Wiley Blackwell, Malden, MA & Oxford.
Lavandera, Beatriz. 1978. Where does the sociolinguistic variable stop? *Language in Society* 7: 171–182.
Mair, Christian. 2006. *Twentieth-century English: History, Variation, and Standardization*, Cambridge University Press, Cambridge.
McEnery, Tony, Xiao, Richard and Tono, Yukio. 2006. *Corpus-based Language Studies: An Advanced Resource Book*, Routledge, New York.
Miller, George A. and Chomsky, Noam. 1963. Finitary models of language users, In Luce, R. D., Bush, R. R. and Galanter, E. (eds.), *Handbook of Mathematical Psychology*, Wiley, New York, pp. 419–491.
Milroy, James, Milroy, Lesley, Hartley, Sue and Walshaw, David. 1994. Glottal stops and Tyneside glottalization: Competing patterns of variation and change in British English. *Language Variation and Change* 6: 327–357.
Montgomery, Chris and Beal, Joan C. 2011. Perpectual dialectology, In Maguire, W. & McMahon, A. (eds.), *Analysing Variation in English*, Cambridge University Press, Cambridge etc., pp. 121–148.
Nerbonne, John. 2009. Data-driven Dialectology. *Language and Linguistics Compass* 3 (1): 175–198.
Nerbonne, John, Heeringa, Wilbert and Kleiweg, Peter. 1999. Edit Distance and Dialect Proximity, In Sankoff, David & Kruskal, Joseph (eds.), *Time*

Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, CSLI Press, Stanford, pp. v–xv.

Nerbonne, John and Wiersma, Wybo. 2006. A Measure of Aggregate Syntactic Distance, In *Proceedings of the Workshop on Linguistic Distances*, pp. 82–90.

Orton, Harold, Sanderson, Steward and Widdowson, John. 1978. *The Linguistic Atlas of England*, Croom Helm, London.

Pietsch, Lukas. 2005. *Variable Grammars: Verbal Agreement in Northern Dialects of English*, Niemeyer, Tübingen.

Pietsch, Lukas. 2012. Verbal concord, In Hickey, R. (ed.), *Areal Features of the Anglophone World*, Mouton de Gruyter, Berlin & New York, pp. 355–378.

Poplack, Shana and Tagliamonte, Sali. 2001. *African American English in the Diaspora*, Blackwell, Oxford.

Rácz, Péter. 2012. Operationalising salience: definite article reduction in the North of England. *English Language and Linguistics* 16: 57–79.

Sampson, Geoffrey. 2002. Regional variation in the English verb qualifier system. *English Language and Linguistics* 6: 17–30.

Séguy, Jean. 1971. La Relation Entre La Distance Spatiale et La Distance Lexicale. *Revue de Linguistique Romane* 35: 335–57.

Spears, Arthur. 1982. The semi-auxiliary *come* in Black-English Vernacular. *Language* 58: 850–872.

Streck, Tobias and Auer, Peter. 2012. Das raumbildende Signal in der Spontansprache. Dialektometrische Untersuchungen zum Alemannischen in Deutschland. *Zeitschrift für Dialektologie und Linguistik* 79(2): 149–188.

Szmrecsanyi, Benedikt. 2006. *Morphosyntactic Persistence in Spoken English: a Corpus Study at the Intersection of Variationist Sociolinguistics, Psycholinguistics, and Discourse Analysis*, Mouton de Gruyter, Berlin/New York.

Szmrecsanyi, Benedikt. 2010. *The morphosyntax of BrE dialects in a corpus-based dialectometrical perspective: feature extraction, coding protocols, projections to geography, summary statistics*. URN: urn:nbn:de:bsz:25-opus-73209, URL: http://www.freidok.uni-freiburg.de/volltexte/7320/. Freiburg. (64pp.)

Szmrecsanyi, Benedikt. 2013. *Grammatical Variation in British English Dialects: A Study in Corpus-Based Dialectometry*, Cambridge University Press, Cambridge.

Szmrecsanyi, Benedikt and Hernández, Nuria. 2007. *Manual of Information to accompany the Freiburg Corpus of English Dialects Sampler*, University of Freiburg, Freiburg. URN:urn:nbn:de:bsz:25-opus-28598, URL: http://www.freidok.uni-freiburg.de/volltexte/2859/.

Szmrecsanyi, Benedikt and Wälchli, Bernhard (eds.). 2014. *Aggregating Dialectology, Typology, and Register Analysis: Linguistic Variation in Text and Speech*, Walter de Gruyter, Berlin.

Tagliamonte, Sali. 1998. *Was/were* variation across the generations: View from the city of York. *Language Variation and Change* 10: 153–191.

Tagliamonte, Sali. 2001. *Come/came* variation in English dialects. *American Speech* 76: 42–61.

Tagliamonte, Sali. 2012a. *Roots of English: Exploring the History of Dialects*, Cambridge University Press, Cambridge.

Tagliamonte, Sali. 2012b. *Variationist Sociolinguistics: Change, Observation, Interpretation*, Wiley Blackwell, Malden, MA & Oxford.

Tagliamonte, Sali and D'Arcy, Alex. 2007. The modals of obligation/necessity in Canadian perspective. *English World-Wide* 28: 47–87.

Tagliamonte, Sali and Smith, Jennifer. 2005. No Momentary Fancy! The Zero 'Complementizer' in English Dialects. *English Language and Linguistics* 9: 289–309.

Tagliamonte, Sali, Smith, Jennifer, and Lawrence, Helen. 2005. 'No Taming the Vernacular!' Insights from the Relatives in Northern Britain. *Language Variation and Change* 17: 75–112.

Trudgill, Peter. 1990. *The Dialects of England*, Blackwell, Cambridge, Mass.

Trudgill, Peter. 2013. *Review of Areal Features of the Anglophone World*, Edited by Raymond Hickey. Berlin & Boston: De Gruyter Mouton, 2012. *Journal of Linguistic Geography* 1: 86–92.

Tubau Muntañá, Susagna. 2008. *Negative Concord in English and Romance: Syntax-Morphology Interface Conditions on the Expression of Negation*, LOT Publications, Utrecht.

Uiboaed, Kristel, Hasselblatt, Cornelius, Lindstrom, Liina, Muischnek, Kadri and Nerbonne, John. 2013. Variation of Verbal Constructions in Estonian Dialects. *Literary and Linguistic Computing* 28 (1): 42–62.

Wolk, Christoph. 2014. *Integrating Aggregational and Probabilistic Approaches to Language Variation*. PhD Dissertation, University of Freiburg.

Zelinsky, Wilbur. 1973. *The Cultural Geography of the United States*, Prentice-Hall, Englewood Cliffs, NJ.