English corpus linguistics: the current state-of-the-art, and a critical appraisal

Benedikt Szmrecsanyi (KU Leuven) and Laura Rosseel (KU Leuven / VUB)

Acknowledgments

Thanks go to those colleagues of ours who participated in our survey and to an anonymous reviewer for helpful suggestions.

1. Introduction

To fix terminology at the outset, we define a corpus as "a collection of texts or parts of texts upon which some general linguistic analysis can be conducted" (Meyer 2002: xi). Corpus linguistics, then, is "a methodology that draws on collections of more or less naturalistic texts or speech for the sake of conducting some sort of linguistic analysis" (Szmrecsanyi 2017: 2). Our point of departure is that the first edition of this handbook had an excellent introduction to corpus linguistics from an English linguistics angle (McEnery & Gabrielatos 2006). But more than a decade later, old debates and dichotomies – e.g. about corpus-based versus corpus-driven approaches (Tognini-Bonelli 2001) – are largely settled, and corpus linguistics has become so uncontroversial and mainstream, at least in English linguistics, that there is a need to re-think the status of corpus linguistics in English linguistics and, also, to re-think the remit of a chapter on corpus linguistics in a Handbook of English linguistics. Why does a Handbook of English Linguistics need a chapter on corpus linguistics at all? Notice that handbooks of the linguistics of other languages do not typically have chapters on corpus linguistics (see e.g. Wetzels, Menuzzi & Costa 2016; Tsujimura 1999; Hualde, Olarrea & O'Rourke 2013; Brown & Yeon 2015). So it seems that corpus-linguistic methods have a special status in English linguistics. Against this backdrop, rather than explaining the corpus-linguistic methodology (for this, we refer the reader to the excellent chapters in Lüdeling & Kytö 2009; Biber & Reppen 2015), the objective of this chapter is to discuss the reasons for the special status of corpus methodologies in English linguistics (or vice versa).

To set the stage, we conducted an informal poll among a convenience sample of *N* = 13 colleagues of ours not specializing in English linguistics. The question that we asked them was the following: "Are there resources (and possibly methods) in English-language corpus linguistics that you wish you had at your disposal but currently don't? Are there ways in which English-language corpus linguistics sets a bad example for best practice in corpus linguistics at large?" The positive feedback that we received may be summarized as follows:

- Eight colleagues found it enviable that data sparseness is typically less of an issue in English-language corpus linguistics compared to other philologies, including in historical English corpus linguistics. Resources that were explicitly mentioned by these colleagues included the Helsinki Corpus, the Corpus of Early English Correspondence, the BYU corpora, MIMIC-III (clinical language), and syntactically annotated corpora such as the Penn Parsed Corpus Series.
- Three colleagues responded that English-language corpus linguists have pioneered rigorous corpus building guidelines, as well as – especially in recent years – a number of innovative data analysis & modeling techniques.

- Two colleagues mentioned that there is a comparatively large number of specialized English-language corpora, covering a wide range of regional varieties, registers, communication types etc.
- Two colleagues stated that good annotation is comparatively common in English-language corpora.
- One colleague found that English-language corpora come with particularly good user interfaces.

Negative feedback included the following:

- Six colleagues complained about a lack of awareness in English corpus linguistics circles that many methods and findings generated in English corpus linguistics (e.g. phraseology, collocation research, *n*-grams) cannot be easily applied to, or are not true for, morphologically rich(er) languages.
- Two colleagues expressed the hope that English corpus linguistics would become more involved in comparative analyses with other languages or at least become more aware of the work that has been done in non-English corpus linguistics and highlight how their work on English contributes to those other language studies.
- One colleague discerned a tendency among some English-language corpus linguists to rest on laurels from the pioneering stages in corpus linguistics, neglecting innovation.
- One colleague noted a certain unhealthy skepticism among English corpus linguists towards theorizing.
- One colleague lamented about a wide-spread focus in English-language corpus linguistics on words as basic linguistic unit.
- One colleague mentioned the focus often being too quantitative with limited attention for the possibilities of qualitative research.

Building on this feedback, the remainder of this chapter is organized as follows. In Section 2, we discuss – in line with responses in our mini-survey – the extent to which English-language corpus linguistics is comparatively well-endowed with resources. Section 3 then moves on to explore why English-language corpus linguistics is seen by many to have created a number of innovative data analysis & modeling techniques. Section 4 concludes the chapter.

2. English language corpora are numerous, large, and well annotated

English-language corpora are simply too numerous to even begin to catalogue them here in any detail. Still, there are a few trends that we should mention. For one thing, in the realm of synchronic English linguistics the field seems to be moving away from the compilation and analysis of "representative" corpora – consider e.g. the 100-million word British National Corpus (BNC) (Aston & Burnard 1998) (http://www.helsinki.fi/varieng/CoRD/corpora/BNC/) – and increasingly turn to specialized corpora such as e.g. the Switchboard Corpus of American English (Godfrey, Holliman & McDaniel 1992), which is widely used to investigate spoken English; the International Corpus of English (ICE) (Greenbaum 1996) (http://www.ice-corpora.uzh.ch/en.html), which is designed to facilitate the investigation of regional varieties of English; dialect corpora such as the Freiburg Corpus of English Dialects (Szmrecsanyi & Hernández 2007) (https://fred.ub.uni-freiburg.de/); the International Corpus of Learner English (ICLE) (Granger, Dagneaux & Meunier 2002) (https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html), which samples essays written by

learners of English; or the corpus of Global Web-Based English (GloWbE) (Davies & Fuchs 2015) (https://corpus.byu.edu/glowbe/), which covers a massive amount of web language.

In historical English linguistics recent years have seen a steady increase in the number of available resources documenting Early and Late Modern English, from the Early English Books Online (EEBO) database (see e.g. https://corpus.byu.edu/eebo/) to A Representative Corpus Of Historical English Registers (ARCHER) (Yáñez-Bouza 2011) (http://www.projects.alc.manchester.ac.uk/archer/) and the 400-million word Corpus of Historical American English (see https://corpus.byu.edu/coha/).

Mair (2006: 355) draws a distinction between "small-and-tidy" corpus linguistics and "big-and-messy" corpus linguistics. Keeping in mind that not all small corpora are necessarily tidy while some big corpora are not messy, the situation in English corpus linguistics at the time of writing is such that the small and tidy tradition is still going strong: data sources such as e.g. the 1-million words Brown corpora (Francis & Kučera 1979; Hinrichs, Smith & Waibel 2010) are still widely used. That being said, it is certainly true that corpora are becoming ever larger. This is a development not at all specific to English-language corpora, but it seems fair to say that the development is particularly pronounced here. Consider the corpora available at https://corpus.byu.edu/: these include e.g. the 6 billion-word News on the Web (NOW) corpus (https://corpus.byu.edu/now/), or the 1.9 billion word Corpus of Global Web-Based English (GloWbE). Needless to say, corpora that are big permit the researcher to tackle research questions, e.g. about lexis, that are impossible to deal with on the basis of smaller corpora. Huge corpora, such as NOW and GloWbE, typically contain written materials that can be fairly easily and automatically harvested from the world-wide web. By contrast, data scarcity is still by and large the name of the game when it comes to corpora covering face-to-face speech, but here too the situation is improving -- consider e.g. the 11.5-million-word spoken component of the British National Corpus 2014 (http://corpora.lancs.ac.uk/bnc2014/), which covers conversations that were gathered from members of the UK public between 2012 and 2016.

Corpus annotation is nowadays fairly uncontroversial and, in fact, widespread in the realm of English corpus linguistics, as was also mentioned by some respondents in our mini-survey. Part-of-speech (POS) annotation is now relatively standard for English-language corpora. Syntax-parsed corpora are increasingly becoming available; consider e.g. the British component of the International Corpus of English (ICE) (http://www.ucl.ac.uk/english-usage/projects/ice-gb/), the Penn Parsed Corpora of Historical English series (http://www.ling.upenn.edu/hist-corpora/), or the English corpora in the Universal Dependencies collection (https://universaldependencies.org/). Lemmatization and synset-annotation is, besides POS-tagging, a feature that for example the BYU corpora come with, in addition to a user interface that makes using these annotation layers fairly easy. Phonetic annotation – consider e.g. the Buckeye Speech Corpus (https://buckeyecorpus.osu.edu/) – is rather rare. The best-annotated corpus in English-language corpus linguistics is probably the Switchboard Corpus, for which multiple annotation layers – e.g. syntax, phonetics, discourse – are available.

3. Some important methodological innovations developed in English-language corpus linguistics

In this section we catalogue seven methodologies pioneered in English-language corpus linguistics. We exclude from the discussion applied corpus linguistics, e.g. for lexicography, pedagogy, or forensic linguistics, and refer the reader instead to the relevant chapters in standard textbooks (e.g. McEnery, Xiao & Tono 2010: 80–122) and handbooks (e.g. Biber & Reppen 2015: Part IV).

3.1. The British tradition in corpus linguistics

The British tradition in corpus linguistics essentially approaches issues relating to meaning and grammar by studying co-occurrence patterns between words. Selected well-known concepts emanating from this research tradition include

- **collocation** ("collocations of a given word are statements of the habitual or customary places of that word"; Firth 1968: 181) – for example, in the Corpus of Contemporary American English the top left collocates of *food* are *fast* (as in *fast food*), *good* (as in *good food*), and *Chinese* (as in *Chinese food*);
- **semantic prosody** (about the "consistent aura of meaning with which a form is imbued by its collocates"; Louw 1993: 157) – for example, Louw (1993: 33–34) shows that in the Cobuild corpus, the right-collocates of *utterly* tend to be unpleasant (as in *utterly terrified*), hence *utterly* has negative semantic prosody.[1]
- **colligation** (about "the grammatical company a word keeps"; Hoey 1997: 8) – for example, Sinclair (Sinclair 1998: 15; discussed in Lehecka 2015) notes that the phrase *naked eye* is often preceded by a preposition and determiner, as in *to the naked eye*.[2]

So it is evident that the British tradition in corpus linguistics takes a particular interest in the intersection of lexis and grammar. The basic idea behind work on "lexicogrammar" (parlance of Halliday 1991; Halliday 1992) or "lexical grammar" (parlance of Sinclair; see e.g. Sinclair 2000) is that a strict separation between lexis and grammar is misguided: "the grammar of a language and its lexicon are not separate entities" (Hunston 2015: 201). Consider Sinclair's (1991: 109–110) distinction between the "slot and filler model" plus the "open choice principle", on the one hand, versus the "idiom principle" on the other hand: according to structuralist thinking, language users can fill slots offered by the grammar with lexical material of their choice. But this will not necessarily yield idiomatic language, hence the idiom principle: idiomatic language use is not a matter of filling at liberty slots afforded by the grammar but rather of respecting collocational preferences. This is another way of saying that being fluent involves using prefabs. This reasoning is why phraseology plays such an important role in the British tradition in corpus linguistics (Hunston 2015: 212).

This line of research has had great impact in English linguistics and beyond, and is often – particularly by linguists working on other languages than English – equated with English corpus linguistics per se. This, then, may also explain some of the less positive comments that shine through in our mini-survey. For one thing, corpus research in the British tradition is typically word-based: in actual practice, what is of key interest is how orthographically transcribed words co-occur with other items or patterns. The advantage of this approach, as noted by McEnery and Gabrielatos (2006: 45), is its practical appeal: it works just fine with raw corpora, and standard corpus analysis software suffices to carry out the analysis, as long as the corpora under analysis are reasonably large. However,

[1] A related notion is that of semantic preference, which is about the "relation between a lemma or word-form and a set of semantically related words" (Stubbs 2001: 111–112).

[2] A recent extension of colligational analysis is that of collostructional analysis (Stefanowitsch & Gries 2003), which uses more advanced statistical machinery to investigate the lexemes that are attracted by particular constructions, rather than the other way round. For example, Stefanowitsch and Gries (2003: 227–230) show that those dative verbs ("collexemes" in their parlance) most strongly attracted to the ditransitive dative construction are *give* (as in *Tom gave me a present*), *tell* (as in *Tom told me a story*), and *send* (as in *Tom send me a letter*). See Gries and Stefanowitsch (2004) and Hilpert (2006) for follow-up work.

outside English corpus linguistics the notion of the orthographically transcribed word is not uncontroversial: Haspelmath and Michaelis (2017: 6), for example, argue that "the notion of "word" cannot be defined consistently across languages (other than orthographically, in languages with spaces between words)". The problem, then, is that if the notion of "word" is questionable cross-linguistically, then so are word-based techniques such as collocation analysis. Still, it is clear that collocation-based techniques work for English and similarly analytic languages (see Xiao 2015: 120–123 for a case study of collocation and semantic prosody in Chinese). That said, it is less clear that word-based techniques developed in the British tradition in corpus linguistics work as well for synthetic and inflectional languages such as e.g. Estonian, where one runs into all sorts of problems having to do with distinguishing between affixation and collocation. This, among other things, is the reason why some respondents in our mini-survey worried that methods developed in English-language corpus linguistics cannot necessarily be applied to morphologically rich(er) languages.

3.2. Corpus-based discourse studies

English-language corpus linguistics is characterized by a fairly recent, but increasingly well-established tradition of research which combines corpus-based methodology with analysis techniques from the field of discourse studies. Within this tradition we count approaches such as corpus-assisted discourse studies (CADS, e.g. Partington, Duguid & Taylor 2013) or certain work from the Critical Discourse Analysis paradigm (CDA, e.g. Baker et al. 2008). These approaches gradually took off around the year 2000 – with a number of precursors such as the work of Stenström (Partington 2004) – and have since gained increasing recognition both on the side of corpus linguistics, as well as within the field of discourse studies.

Researchers working on the intersection of these fields are mainly interested in the study of various aspects of discourse (e.g. evaluation or discourse organization), but use quantitative methods from corpus linguistics, such as collocation analysis, to complement the typically qualitative tools traditionally used in discourse studies. In that sense, this branch of linguistics builds on the tradition of Lexical Grammar presented above.

While discourse studies typically work with smaller datasets, the tradition orienting towards corpus linguistics uses larger amounts of data for their analyses. These large corpora are often used in the first exploratory phase of a study using concepts like collocations and semantic prosody to identify patterns, topics, and topoi in text (Gabrielatos 2008). Such an exploratory analysis then serves as a starting point for a subsequent, more in-depth qualitative study of the texts. To illustrate this approach, let us take a look at Gabrielatos & Baker (2008) who used a 140-million word corpus of British news articles focusing on migration to study the representation of refugees and asylum seekers in the British press. They retrieve the collocates of terms referring to people in these positions and use those as 'a clear indication of the stance of the writer/newspaper toward these groups' (Gabrielatos & Baker 2008: 14).

Another way in which corpora are used in this branch of discourse studies is as a tool for tracking down and selecting texts of interest for a discursive analysis in a consistent and transparent way (e.g. Forchtner & Kølvraa 2012). Like in the example from Gabrielatos & Baker (2008) above, this approach uses corpora as a first step towards a discursive study. Still another approach to combine corpora with traditional approaches within discourse studies is to use them as an additional validation of the results obtained in the initial analysis which is typically based on a smaller dataset. Corpora can provide additional data which allow to corroborate and generalize evidence from a discourse analytic study on a small sample. In this way, combining quantitative corpus methods and qualitative discourse tools in

a dialectic procedure where the results of one analysis inspire the other and vice versa, allows discourse analysts to arrive at a more nuanced understanding of the phenomenon under study, but also to explore the pervasiveness of the discourse patterns under study in larger collections of data (Jaworska 2016).

3.3. Corpus-based approaches to dialectology and regional varieties

Dialectology and research into geographical/regional variation are needless to say old and mature research fields in linguistics. Traditional dialectology is concerned with 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' (Trudgill 1990: 5). The traditional data source in dialectology, besides anecdotal evidence, are dialect atlases, such as e.g. the Survey of English Dialects (SED) (Orton & Dieth 1962). Starting in the late 1990s, however, dialectologists have begun to turn to dialect corpora, which tend to cover orthographically transcribed interviews with dialect speakers, similar to sociolinguistic interviews (see Szmrecsanyi & Anderwald 2017: 301 for discussion). The emergence of dialect corpora has put frequency and intra-speaker variation on the map in dialectology (as opposed to the more structural, categorical information available in dialect atlases), and the trend towards usage of dialect corpora has been particularly pronounced in English dialectology. Early studies include Anderwald (2003), who studies non-standard negation patterns (as in *he don't have money*) based on the spoken material in the British National Corpus, which is (partly) annotated for the regional provenance of the speakers; Tagliamonte and Smith (2002), who investigate variation in NEG/AUX contraction (as in *it isn't true* versus *it's not true*) in corpora covering vernacular speech coming from eight communities in the UK; the papers in Kortmann et al. (2005), which investigate non-standard grammar based on the Freiburg Corpus of English Dialects (FRED), which contains interviews with dialect speakers from all over England, Wales, and Scotland; and Beal and Corrigan (2006), who study negation in Tyneside English, drawing on the Newcastle Electronic Corpus of Tyneside English (NECTE). More recent methodological innovations include corpus-based dialectometry, which aggregates over frequencies of many features to calculate measures of dialect distance and similarity as a function of geographic distance (see e.g. Wolk & Szmrecsanyi 2018), and usage of more unorthodox dialectology corpora, such as the corpus of letters to the editor from all over the U.S. analyzed in Grieve (2016).

Beyond dialectology with its focus on (more or less) traditional dialects, the English language of course offers exciting opportunities to study differences between a vast number of regional varieties of English around the world, thanks to colonial activities of the British Empire. A classic topic of the literature in this connection are differences between American and British English (see e.g. the papers in Rohdenburg & Schlüter 2009). Earlier research on British-American differences has profited enormously from the compilation of the Brown family corpora, which are matching 1-million word corpora of standard written-edited English and which facilitate not only the contrastive investigation of the two standard varieties in their written form, but also the study of language change in progress (see e.g. Hundt & Mair 1999 for seminal work). This rich literature is unmatched in other philologies.

English corpus linguistics is leading when it comes to study of post-colonial varieties around the world, including not only what Kachru (1992) has called "Inner Circle" varieties such as American English, British English, New Zealand English and so on but also a large number of "Outer Circle" varieties such as e.g. Indian English, Singapore English, or Nigerian English. Not all of the research on post-colonial varieties is corpus-based; surveys such as the electronic World Atlas of Varieties of English (Kortmann & Lunkenheimer) play an important role. But the emergence of World Englishes

corpora has clearly boosted the field. Consider the International Corpus of English (ICE) series, whose goal it is to compile 1-million word matching corpora to document Inner Circle or Outer Circle varieties of English around the world (Greenbaum 1996), or the Corpus of Global Web-based English (GloWbE), whose approximately 2 billion words of running text cover some 20 English-speaking countries around the world (Davies & Fuchs 2015). Recent representative studies using these corpora include Tamaredo (2018), who taps into ICE-India, ICE-Singapore and ICE-GB to study pronoun omission (as in ___ *can't say I like it*), and Schmidtke and Kuperman (2017), who tap into the GloWbE corpus to study noun countability (as in *two luggages*) in World Englishes. Other languages with a similar post-colonial reach, such as French and Spanish, are not nearly as well documented with corpora.

3.4. Multidimensional Analysis

Multidimensional analysis is a quantitative corpus-based approach within the field of register studies. The approach focuses on linguistic variation that is determined by situational variables: the distribution of linguistic features in a text heavily depends on the register it belongs to (e.g. newspapers, academic lectures, sales pitches, application letters). Although the importance of register for linguistic variation has been recognized and studied for a long time, it was not until the 1980s with the work of Biber (especially Biber 1986; ibid. 1988) that register variation was studied quantitatively using large corpora. What sets the work of Biber and his collaborators apart from previous work on register is that is does not focus on the behavior of a single linguistic feature in various types of texts, but that it investigates the co-occurrence patterns of many different linguistic features in various registers. This is done through multidimensional analysis (MDA), a technique that identifies groupings of linguistic features by using factor analysis (Biber 1986; ibid. 1988). The dimensions (i.e. clusterings of features) resulting from that analysis are then interpreted according to their communicative function (Conrad 2015: 316-317). The next step in the analysis is then to compare how various registers score on these dimensions or how much variability the registers show on each of the dimensions that were identified. Hence, MDA does not only allow to study how multiple feature vary across register, it also allows to characterize registers based on the co-occurrence patters of linguistic features.

To make this more concrete, let us briefly consider a concrete example from Biber's seminal 1986 article which compares spoken and written language based on 41 linguistic features in over 500 text samples. One of the three dimensions that Biber found was characterizing the distinction between spoken and written texts showed a clustering of features like word length (i.e. longer vs. shorter words) and type/token ratio (i.e. how much variation there is in word choice) (Biber 1986: 394). Interpreting the communicative functions of these features, the analysis shows that what these texts have in common is a high density of precise information, which is typical of written texts which are often (extensively) edited. Other features which appear on the same dimension are the use of *yes*/*no*-questions, *wh*-questions, and the pronouns *I* and *you*. These are interpreted as marking direct interaction which is typical of spoken language. As a result, Biber labels this dimension 'Interactive vs. Edited text' (Biber 1986: 395).

MDA has been used to study a wide range of registers, both from a synchronic and diachronic perspective (Conrad 2015). The large body of MDA research abundantly demonstrates and documents the importance of situational factors for linguistic variation. Yet, the accomplishments of MDA studies go beyond that, as they offer the potential for practical applications like the description of registers and development of study materials for L2 learners, as well as a better informed selection of text genres in the compilation of corpora (Conrad 2015: 317).

From the very beginning, MDA approaches to register studies have focused predominantly on register variation in English. One of the reasons for this, as also pointed out in previous paragraphs, may well be the availability of various large English language corpora that encompass a wide variety of text types. A subfield of English Studies that is particularly associated with MDA studies is English for Specific Purposes (Conrad 2015: 318-319). An example of such work is Rooy & Terblanche (2006) who use MDA to compare native and learner writing regarding aspects of involvement. Building on Biber (1988), the study furthermore compares student writing to other registers like academic writing and spoken language. This strong focus of MDA research on English and English for specific purposes does not mean, however, that this type of research has not developed in branches of linguistics focusing on languages other than English (e.g. Biber et al. 2008 on Spanish; Asención-Delaney & Collentine 2011 on L2 Spanish; Biber & Hared 1992 on Somali). Yet the lion's share of MDA work is still concerned with English and in that respect it is telling that in a chapter summarizing register studies Conrad (2015) includes the application of MDA to languages other than English in her discussion of the extensions of MDA. This shows that work on English serves as a reference point in this line of research.

3.5. Corpus-based psycholinguistics

Psycholinguistics is the field of study that is concerned with how language users produce and process language. The use of corpora in this field takes two forms, both of which are characterized by a rather dominant position of English corpus linguistics. First, we briefly discuss the branch of psycholinguistic research that uses corpus linguistic methods to gain insight into mechanisms of language processing and aims to complement the experimental work typically conducted in psycholinguistics. Next, we will touch upon the indirect use of corpora in psycholinguistics as a resource for materials to design, construct and analyze experiments.

Although most work on the cognitive processes behind speech production and perception builds on a wide variety of experimental paradigms, some corpus studies have been carried out in this domain as well. These corpus studies have their roots in various disciplines ranging from traditional corpus linguistics to computational linguistics and cognitive science. Although experimental studies offer the advantage of carefully controlling for confounding factors in the psycholinguistic processes under study, corpus studies have the benefit of using naturalistic data and allow the possibility to take into account more factors than can be controlled for in experimental settings (Gries 2005).

One example of a topic in psycholinguistics that has received a fair amount of corpus-based attention is priming. Priming refers to the phenomenon that speakers have a tendency to repeat the same linguistic elements (Pickering & Ferreira 2008). In the case of syntactic priming, for instance, a syntactic structure has a higher chance of being used if it was previously used in discourse. A study that looked at priming using traditional corpus linguistic methods of regression modelling is Gries (2005). Focusing on two syntactic alternations in English, i.e. the dative alternation and the particle placement alternation, the author presents corpus-based evidence for earlier accounts of syntactic priming based on experimental research. Dubey, Keller & Sturt (2008) by contrast approach a similar topic from a computational perspective. They present a study that investigates the mechanism underlying the repetition of coordinate structures in discourse using corpora to train computational models. Building on these computer simulations, the authors compare the predictions of two competing accounts for the observed repetitions based on previous experimental research. Aside from priming research, another topic relating to language processing that has been studied rather extensively from a corpus-based perspective is surprisal and informativity. This type of work is exemplified by the studies on the role of predictability of phonological phenomena such as Jurafsky et al. (2001), Bell et al. (2003),

Demberg et al. (2012) on reduction, or Cohen Priva (2017) on lenition processes. More recently, the role of predictability of syntactic variation (e.g. presence or deletion of complementizer *that*) has also received considerable scholarly attention by for instance Jaeger and colleagues (e.g. Jaeger 2010) who have used corpora to study the hypothesis that speakers tend to spread new information evenly across utterances.

In addition to the type of corpus-based work that aims to directly contribute to the understanding of language processing by supplying complementary evidence to experimental studies, corpora are also used to other ends in psycholinguistics. More particularly, they can provide crucial information to build well-controlled experiments or are used to extract naturalistic experimental stimuli (but see Mandera, Keuleers & Brysbaert 2015 for a discussion of limitations of corpus-based methods in this regard). One clear example here is the use of subtitle corpora to extract frequency information used in myriads of psycholinguistic studies to approximate the frequency of words in interaction (New et al. 2007). Additionally, corpora are used to extract probabilities used in computational models of various aspects of language processing (see Roland & Hare 2012 for a discussion of such use of corpora in sentence comprehension research).

3.6. Variationist (socio)linguistics

Variationist sociolinguistics is a vibrant research field whose primary mission is to understand the factors regulating the ways in which language users choose between "alternate ways of saying 'the same' thing" (Labov 1972: 188), with a particular interest in how social factors constrain choices. Consider e.g. Weiner and Labov (1983): the study investigates, on the basis of transcribed spontaneous material from interviews with working-class white speakers in Philadelphia, variation between agentless passives (as in *The liquor closet got broken into*) and "empty" actives (as in *They broke into the liquor closet*) (exemplification from Weiner & Labov 1983: 34). The study investigates a range of factors potentially constraining the choice, and concludes that the single most powerful factor to influence the choice of actives vs. passives is repetition of previous structure. Now, variationist sociolinguistics is a powerful research paradigm that has pioneered a number of innovative analysis techniques – for example, since the 1970s variationist sociolinguistics have been using regression analysis to analyze variation data, in the form of the Varbrul program (see Cedergren & Sankoff 1974). The reason why we mention variationist sociolinguistics in this article is that much of the work carried out in this framework is concerned with English, although (Canadian) French and (North American) Spanish are also going strong. The big question is, however, whether orthodox variationist linguistics qualifies as "corpus-based". To the extent that work in variationist sociolinguistics is based on the analysis of fully transcribed sociolinguistic interviews (the preferred data type in the field), variationist sociolinguistics is arguably corpus-based (see Szmrecsanyi 2017 for discussion). But then again, variationist sociolinguists do not regularly self-identify as corpus linguistics, while some card-carrying corpus linguists feel uneasy about attaching the "corpus" label to transcribed sociolinguistic interviews. What is more, the transcribed interviews are often not publicly accessible.

Even so, work in variationist sociolinguistics has inspired – or is at the very least methodologically allied to – a good many variationist studies in the recent literature that are working on the basis of "proper" and publicly accessible corpora. This line of research may be labeled "corpus-based variationist linguistics" (CVL) (Szmrecsanyi 2017: 3), and includes corpus-based research that meets the following criteria:

1. CVL analysts properly define variables and variants to study different ways of saying the same thing (Labov 1972: 188).
2. Therefore, CVL analysts observe the Principle of Accountability (Labov 1969: 738) and focus on choice-making processes rather than on text frequencies (see Biber et al. 2016 for discussion).
1. CVL uses rigorous quantitative methodologies and statistical modeling techniques (see Tagliamonte & Baayen 2012 for an overview).

Some recent representative studies that come under the remit of CVL research include the following: Bresnan et al. (2007) investigate variation between interchangeable observations of the ditransitive dative construction (as in *Tom sent the president a letter*) and the prepositional dative construction (as in *Tom sent a letter to the president*) based on the Switchboard Corpus of American English (Godfrey, Holliman & McDaniel 1992) and the Treebank Wall Street Journal collection of news and financial reportage (https://catalog.ldc.upenn.edu/LDC2015T13) and via regression modeling demonstrate, among other things, that dative variation is regulated by about ten language-internal/contextual probabilistic constraints; Gries and Hilpert (2010) conduct a regression analysis of variation between 3rd person singular inflections (as in *he giveth* versus *he gives*) in the Parsed Corpus of Early English Correspondence (PCEEC) (http://www-users.york.ac.uk/~lang22/PCEEC-manual/), showing that the change from *–(e)th* to *–(e)s* consisted of five stages; and Hinrichs et al. (2015) model, again via regression analysis, variation between interchangeable restrictive relativizers (*the house which I bought* versus *the house that I bought* versus *the house ___ I bought*) in the Brown family of corpora (Hinrichs, Smith & Waibel 2010). Analysis shows that the shift from restrictive *which* to restrictive *that* in late 20th century English is best characterized as a case of institutionally backed colloquialization.

Again, the point is that many if not most CVL studies investigate variation in English, although some other languages (e.g. Dutch -- see, for example, Grondelaers & Speelman 2007; Levshina, Geeraerts & Speelman 2013; Pijpops & Van de Velde 2014) are also going strong.

3.7. Learner corpus research

Learner corpus research (LCR) is a fairly young research endeavor that started in the late 1980s as a movement to "revolutionize" (Granger 1994) applied linguistics by marrying second language acquisition (SLA) research to corpus linguistics. In contrast to traditional SLA, LCR emphasizes performance rather than competence: what takes center stage is frequency, collocations, lexicogrammar, and message conveyance (see Gilquin & Granger 2015: 418–420 for discussion). The corpora investigated in LCR have traditionally covered primarily written production in aggregated learner populations; consider e.g. the International Corpus of Learner English (ICLE) (Granger et al. 2009), which contains writing by higher intermediate to advanced learners of English from numerous mother tongue backgrounds. Increasingly, however, spoken LCR resources are coming online, such as the Louvain International Database of Spoken English Interlanguage (LINDSEI; Gilquin et al. 2010), which contains interviews in English with university students from several mother tongue backgrounds and whose English proficiency ranges from intermediate to advanced.

Some representative studies include the following. Gilquin and Paquot (2008) investigate how learners of English perform twelve rhetorical functions in academic English, in comparison to native academic English. The study checks the frequencies of these features in three (sub)corpora: native academic essays covered in the British National Corpus (BNC), spoken materials covered in the BNC, and non-native academic writing as sampled in ICLE. Analysis shows that compared to native

speakers learners overuse spoken-like features in their academic writing, indicating that they are not sufficiently aware of register differences. For example, as far as the expression of possibility is concerned, *perhaps* and *maybe* are lexical variants, but while in native academic writing *perhaps* is vastly more frequent than *maybe*, in non-native academic writing we find that *maybe* is used approximately as frequently as *perhaps*, which resembles the rates that we find in the spoken sections of the BNC (Gilquin & Paquot 2008: Figure 1).

Gries and Deshors (2014) (see also Gries & Deshors 2015) is a paper that illustrates the recent trend in LCR towards using more advanced statistical methods. The authors use a methodology – Multifactorial Prediction and Deviation Analysis with Regressions (MuPDAR for short) – which is basically a two-step variationist regression analysis procedure. MuPDAR specifically compares native to non-native performance in scenarios where language users have the choice between different ways of saying the same thing. Faced with such a choice, MuPDAR provides a model of learners' choices given a range of contextual factors while asking what a native speaker would do under the same circumstances. As a case study, Gries and Deshors explore variation between *can* and *may* (as in *we can also let our imagination wander* versus *we may also let our imagination wander*) in the French and Chinese subsections of ICLE, as well as in the Louvain Corpus of Native English Essays (LOCNESS) as a native benchmark.

Ehret and Szmrecsanyi (in press), on the other hand, is a study that demonstrates how LCR is also increasingly opening up to debates in general linguistics at large. SLA analysts have long been concerned with how to measure the complexity of interlanguages, but the point of departure of Ehret and Szmrecsanyi (in press) is the fact that since the early 2000s, cross-linguistic typologists and sociolinguists have increasingly experimented with new and innovative ways to measure language complexity. The paper "imports", as it were, one such complexity measure into LCR: drawing on information theory, the study defines the complexity of a text as proportional to the length of the shortest algorithm that can generate that text (Kolmogorov 1963; Kolmogorov 1965). With this construct under its belt, the study assesses the complexity of learner essays sampled in ICLE. Analysis shows that, among other things, increased L2 instructional exposure predicts increased linguistic complexity of the essay material.

The reason why LCR deserves its own section in this article is that while LCR is now a mature research field with its own journal (the *International Journal of Learner Corpus Research*), and its own association (https://www.learnercorpusassociation.org/). Most of the pioneering work in this tradition has been done by English linguists on learner English (Gilquin & Granger 2015: 428), thanks in no small part to the early availability of exquisite English-language resources such as ICLE.

4. Conclusion

Corpus-based research in English linguistics is by now so common that it would be an utterly hopeless task to even begin to summarize this literature comprehensively in a single article. Instead, we opted to reflect on the status of corpus-linguistic methodologies in English linguistics, and on the role of English linguistics in the development of corpus linguistics: what does English-language corpus linguistics look like from the outside? What is the extent to which English-language corpus linguistics is comparatively well-endowed with resources, in a way that other languages are not? And finally, what are key corpus-linguistic approaches and methodologies that were mainly or entirely developed in the context of English linguistics? In connection with that last question, we then sketched seven corpus-linguistic approaches and methodologies that have (or had initially) a strong English-linguistics bent: The British tradition in corpus linguistics, critical discourse analysis, corpus-based

approaches to dialectology and regional varieties, multidimensional analysis, corpus-based psycholinguistics, variationist linguistics, and learner corpus research.

References

Anderwald, Lieselotte. 2003. Non-standard English and typological principles: The case of negation. In Günter Rohdenburg & Britta Mondorf (eds.), *Determinants of Linguistic Variation*, 507–529. Berlin, New York: Mouton de Gruyter.

Asención-Delaney, Yuly & Joseph Collentine. 2011. A Multidimensional Analysis of a Written L2 Spanish Corpus. *Applied Linguistics* 32(3). 299–322.

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

Baker, Paul, Costas Gabrielatos, Majid Khosravinik, Michal Krzyzanowski, Tony McEnery & Ruth Wodak. 2008. A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society* 19(3). 273–306. doi:10.1177/0957926508088962.

Beal, Joan C & Karen P Corrigan. 2006. No, Nay, Never: Negation in Tyneside English. In Yoko Iyeiri (ed.), *Aspects of English Negation*, 139–157. Amsterdam: Benjamins.

Bell, Alan, Daniel Jurafsky, Eric Fosler-Lussier, Cynthia Girand, Michelle Gregory & Daniel Gildea. 2003. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America* 113(2). 1001–1024. doi:https://doi.org/10.1121/1.1534836.

Biber, Douglas. 1986. Spoken and written textual dimensions in English: Resolving the contradictoriy findings. *Language* 62. 384–414.

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

Biber, Douglas, Mark Davies, James K. Jones & Nicole Tracy-Ventura. 2008. Spoken and written register variation in Spanish: A multi-dimensional analysis. *Corpora* 1(1). 1–37.

Biber, Douglas, Jesse Egbert, Bethany Gray, Rahel Oppliger & Benedikt Szmrecsanyi. 2016. Variationist versus text-linguistic approaches to grammatical change in English: Nominal modifiers of head nouns. In Merja Kytö & Päivi Pahta (eds.), *The Cambridge Handbook of English Historical Linguistics* (Cambridge Handbooks in Language and Linguistics), 351–375. Cambridge: Cambridge University Press.

Biber, Douglas & Mohamed Hared. 1992. Dimensions of register variation in Somali. *Language Variation and Change* 4(1). 41–75.

Biber, Douglas & Randi Reppen (eds.). 2015. *The Cambridge handbook of English corpus linguistics* (Cambridge Handbooks in Language and Linguistics). Cambridge: Cambridge University Press.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina & Harald Baayen. 2007. Predicting the Dative Alternation. In Gerlof Boume, Irene Krämer & Joost Zwarts (eds.), *Cognitive Foundations of Interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Brown, Lucien & Jaehoon Yeon. 2015. *The Handbook of Korean Linguistics*. New York, NY: John Wiley & Sons. http://nbn-resolving.de/urn:nbn:de:101:1-201506101078 (23 March, 2018).

Cedergren, Henrietta & David Sankoff. 1974. Variable rules: Performance as a statistical reflection of competence. *Language* 50(2). 333–355.

Cohen Priva, Uriel. 2017. Informativity and the actuation of lenition. *Language* 930(3). 569–597. doi:doi: 10.1353/lan.2017.0037.

Conrad, Susan. 2015. Register variation. In Douglas Biber & Randi Reppen (eds.), *The Cambridge handbook of English corpus linguistics*. Cambridge: Cambridge University Press.

Davies, Mark & Robert Fuchs. 2015. Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide* 36(1). 1–28.

Demberg, Vera, Asad B. Sayeed, Philip J. Gorinski & Nikolaos Engonopoulos. 2012. Syntactic surprisal affects spoken word duration in conversational contexts. *Proceedings of the 2012 Joint*

Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning 356–367.
Dubey, Amit, Frank Keller & Patrick Sturt. 2008. A Probabilistic Corpus-based Model of Syntactic Parallelism. *Cognition* 109(3). 326–344.
Ehret, Katharina & Benedikt Szmrecsanyi. in press. Compressing learner language: An information-theoretic measure of complexity in SLA production data. *Second Language Research*. doi:10.1177/0267658316669559. http://slr.sagepub.com/cgi/doi/10.1177/0267658316669559 (17 October, 2016).
Firth, John. 1968. A synopsis of linguistic theory, 1930-1955. In F. Palmer (ed.), *Selected Papers of J. R. Firth 1952-59*, 168–205. London: Longmans.
Forchtner, Brenhard & Christoffer Kølvraa. 2012. Narrating a 'new Europe': From 'bitter past' to self-righteousness? *Discourse & Society* 23. 377–240. doi:10.1177/0957926512441108.
Francis, Winthrop Nelson & Henry Kučera. 1979. *Manual of information to accompany a standard corpus of present-day edited American English, for use with digital computers*. From \textlesshttp://khnt.hit.uib.no/icame/manuals/brown/\textgreater.
Gabrielatos, Costas. 2008. Collocational analysis as a gateway to critical discourse analysisThe case of the construction of refugees, asylum seekers and immigrants in the UK press. English Language Institute, University of Michigan.
Gabrielatos, Costas & Paul Baker. 2008. Fleeing, Sneaking, Flooding. A Corpus Analysis of Discursive Constructions of Refugees and Asylum Seekers in the UK Press, 1996-2005. *Journal of English Linguistics* 36(1). 5–38. doi:10.1177/0075424207311247.
Gilquin, Gaëtanelle & Sylviane Granger. 2015. Learner language. In Douglas Biber & Randi Reppen (eds.), *The Cambridge Handbook of English Corpus Linguistics*, 418–436. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139764377.024. https://www.cambridge.org/core/product/identifier/9781139764377%23CN-bp-23/type/book_part (5 October, 2018).
Gilquin, Gaëtanelle & Magali Paquot. 2008. Too chatty: Learner academic writing and register variation. *English Text Construction* 1(1). 41–61. doi:10.1075/etc.1.1.05gil.
Godfrey, John J, Edward C Holliman & Jane McDaniel. 1992. SWITCHBOARD: telephone speech corpus for research and development. *IEEE International Conference on Speech, and Signal Processing, ICASSP-92*, vol. 1, 517–520.
Granger, Sylviane. 1994. The Learner Corpus: a revolution in applied linguistics. *English Today* 10(03). 25. doi:10.1017/S0266078400007665.
Granger, Sylviane, Estelle Dagneaux & Fanny Meunier (eds.). 2002. *The International Corpus of Learner English: Handbook and CD-ROM*. Louvain-la-Neuve: Presses universitaires de Louvain.
Granger, Sylviane, Estelle Dagneaux, Fanny Meunier & Magali Paquot. 2009. *The International Corpus of Learner English: Handbook and CD-ROM*. Louvain-la-Neuve: Presses universitaires de Louvain.
Greenbaum, Sidney. 1996. *Comparing English worldwide: the International Corpus of English*. Oxford, New York: Clarendon Press.
Gries, S. T & A. Stefanowitsch. 2004. Extending collostructional analysis: A corpus-based perspective on alternations. *International Journal of Corpus Linguistics* 9(1). 97–129.
Gries, Stefan Th. 2005. Syntactic Priming: A Corpus-based Approach. *Journal of Psycholinguistic Research* 34(4). 365–399.
Gries, Stefan Th. & Sandra C. Deshors. 2014. Using regressions to explore deviations between corpus data and a standard/target: two suggestions. *Corpora* 9(1). 109–136. doi:10.3366/cor.2014.0053.
Gries, Stefan Th. & Sandra C. Deshors. 2015. EFL and/vs. ESL?: A multi-level regression modeling perspective on bridging the paradigm gap. *International Journal of Learner Corpus Research* 1(1). 130–159. doi:10.1075/ijlcr.1.1.05gri.

Gries, Stefan Th. & Martin Hilpert. 2010. Modeling diachronic change in the third person singular: a multifactorial, verb- and author-specific exploratory approach. *English Language and Linguistics* 14(03). 293–320. doi:10.1017/S1360674310000092.

Grieve, Jack. 2016. *Regional variation in written American English* (Studies in English Language). Cambridge: Cambridge University Press.

Grondelaers, Stefan & Dirk Speelman. 2007. A variationist account of constituent ordering in presentative sentences in Belgian Dutch. *Corpus Linguistics and Linguistic Theory* 3(2). doi:10.1515/CLLT.2007.010. http://www.degruyter.com/view/j/cllt.2007.3.issue-2/cllt.2007.010/cllt.2007.010.xml (7 September, 2015).

Halliday, Michael AK. 1991. Corpus studies and probabilistic grammar. In Karin Aijmer & Bengt Altenberg (eds.), *English corpus linguistics: studies in honour of Jan Svartvik*, 30–40. London: Longman.

Halliday, Michael AK. 1992. Language as system and language as instance: The corpus as a theoretical construct. *Directions in corpus linguistics: proceedings of the Nobel Symposium 82, Stockholm, 4–8 August 1991*, 61–77. Berlin: Mouton de Gruyter.

Haspelmath, Martin & Susanne Maria Michaelis. 2017. Analytic and synthetic: Typological change in varieties of European languages. In Isabelle Buchstaller & Beat Siebenhaar (eds.), *Studies in Language Variation*, vol. 19, 3–22. Amsterdam: John Benjamins Publishing Company. doi:10.1075/silv.19.01has. https://benjamins.com/catalog/silv.19.01has (4 October, 2018).

Hilpert, Martin. 2006. Distinctive collexeme analysis and diachrony. *Corpus Linguistics and Linguistic Theory* 2(2). doi:10.1515/CLLT.2006.012. http://www.degruyter.com/view/j/cllt.2006.2.issue-2/cllt.2006.012/cllt.2006.012.xml (26 November, 2012).

Hinrichs, Lars, Nicholas Smith & Birgit Waibel. 2010. Manual of information for the part-of-speech-tagged, post-edited "Brown" corpora. *ICAME Journal* 34. 189–231.

Hinrichs, Lars, Benedikt Szmrecsanyi & Axel Bohmann. 2015. Which-hunting and the Standard English relative clause. *Language* 91(4). 806–836. doi:10.1353/lan.2015.0062.

Hoey, Michael. 1997. From concordance to text structure: new uses for computer corpora. In B. Lewandowska-Tomaszczyk & P.J. Melia (eds.), *Practical Applications in Language Corpora (PALC '97)*, 2–23. çódå: çódå University Press.

Hualde, José Ignacio, Antxon Olarrea & Erin O'Rourke. 2013. *The Handbook of Hispanic Linguistics*. New York, NY: John Wiley & Sons. http://nbn-resolving.de/urn:nbn:de:101:1-201502032688 (23 March, 2018).

Hundt, Marianne & Christian Mair. 1999. `Agile' and `uptight' genres: the corpus-based approach to language change in progress. *International Journal of Corpus Linguistics* 4. 221–242.

Hunston, Susan. 2015. Lexical grammar. In Douglas Biber & Randi Reppen (eds.), *The Cambridge Handbook of English Corpus Linguistics*, 201–215. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139764377.012. https://www.cambridge.org/core/product/identifier/9781139764377%23CN-bp-11/type/book_part (1 October, 2018).

Jaeger, T. Florian. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology* 61. 23–62.

Jaworska, Sylvia. 2016. Using a corpus-assisted discourse studies (CADS) approach to investigate constructions of identities in media reporting surrounding mega sport events: the case of the London Olympics 2012. In R. I. Lamond & L. Platt (eds.), *Critical Event Studies. Leisure Studies in a Global Era*, 149–174. London: Palgrave Macmillan.

Jurafsky, Daniel, Alan Bell, Michelle Gregory & William D. Raymond. 2001. Probabilistic Relations between Words: Evidence from Reduction in Lexical Production. In Joan Bybee & Paul Hopper (eds.), *Frequency and the emergence of linguistic structure*, 229–254. Amsterdam: John Benjamins Publishing Company.

Kachru, Braj B. (ed.). 1992. *The Other tongue: English across cultures* (English in the Global Context). 2nd ed. Urbana: University of Illinois Press.

Kolmogorov, Andrej. 1963. On Tables of Random Numbers. *Sankhya* (A) 25. 369–375.
Kolmogorov, Andrej N. 1965. Three Approaches to the Quantitative Definition of Information. *Problemy Peredachi Informatsii* 1(1). 3–11.
Kortmann, Bernd, Tanja Herrmann, Lukas Pietsch & Susanne Wagner. 2005. *A comparative grammar of British English dialects agreement, gender, relative clauses*. Berlin; New York: Mouton de Gruyter. http://site.ebrary.com/id/10197230 (6 January, 2014).
Kortmann, Bernd & Kerstin Lunkenheimer (eds.). *The Electronic World Atlas of Varieties of English.* Lepzig: Max Planck Institute for Evolutionary Anthropology. http://ewave-atlas.org.
Labov, William. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45. 715–762.
Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Philadelphia Press.
Lehecka, Tomas. 2015. Collocation and colligation. In Jan-Ola Östman & Jef Verschueren (eds.), *Handbook of Pragmatics*. Amsterdam: John Benjamins Publishing Company. doi:10.1075/hop.19.col2. https://benjamins.com/online/hop/articles/col2 (4 October, 2018).
Levshina, Natalia, Dirk Geeraerts & Dirk Speelman. 2013. Towards a 3D-grammar: Interaction of linguistic and extralinguistic factors in the use of Dutch causative constructions. *Journal of Pragmatics* 52. 34–48. doi:10.1016/j.pragma.2012.12.013.
Louw, Bill. 1993. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis & Elena Tognini-Bonelli (eds.), *Text and technology: In honour of John Sinclair*, 157–176. Amsterdam: Benjamins.
Lüdeling, Anke & Merja Kytö (eds.). 2009. *Corpus linguistics: an international handbook* (Handbücher Zur Sprach- Und Kommunikationswissenschaft = Handbooks of Linguistics and Communication Science Bd.-29.2). Berlin ; New York: Walter de Gruyter.
Mair, Christian. 2006. Tracking ongoing grammatical change and recent diversification in present-day standard English: the complementary role of small and large corpora. In Antoinette Renouf & Andrews Kehoe (eds.), *The Changing Face of Corpus Linguistics: Papers from the 24th International Conference on English Language Research on Computerized Corpora (ICAME 24)*, 355–376. Amsterdam: Rodopi.
Mandera, Paweł, Emmanuel Keuleers & Marc Brysbaert. 2015. How useful are corpus-based methods for extrapolating psycholinguistic variables. *The Quarterly Journal of Experimental Psychology* 68(8). 1623–1642.
McEnery, Tony & Costas Gabrielatos. 2006. English corpus linguistics. *The handbook of English linguistics*, 33–71. Malden, MA ; Oxford: Blackwell Pub.
McEnery, Tony, Richard Xiao & Yukio Tono. 2010. *Corpus-based language studies: an advanced resource book* (Routledge Applied Linguistics). Reprinted. London: Routledge.
Meyer, Charles F. 2002. *English corpus linguistics: an introduction* (Studies in English Language). Cambridge, UK ; New York: Cambridge University Press.
New, Boris, Marc Brysbaert, Jean Veronis & Christophe Pallier. 2007. The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics* 28(4). 661–677.
Orton, Harold & Eugen Dieth. 1962. *Survey of English Dialects*. Leeds: E. J. Arnold.
Partington, Alan. 2004. Corpora and discourse, a most congruous beast. *Corpora and Discourse*, 9–18. Bern: Peter Lang.
Partington, Alan, Alison Duguid & Charlotte Taylor. 2013. *Patterns and Meanings in Discourse: Theory and practice in corpus-assisted discourse studies (CADS)*. Amsterdam/Philadelphia: John Benjamins Publishing Company.
Pickering, Martin J. & Victor S. Ferreira. 2008. Structural priming: A critical review. *Psychological Bulletin* 134(3). 427–459.
Pijpops, Dirk & Freek Van de Velde. 2014. A multivariate analysis of the partitive genitive in Dutch. Bringing quantitative data into a theoretical discussion. *Corpus Linguistics and Linguistic Theory* 0(0). doi:10.1515/cllt-2013-0027. http://www.degruyter.com/view/j/cllt.ahead-of-print/cllt-2013-0027/cllt-2013-0027.xml (17 September, 2015).

Rohdenburg, Günter & Julia Schlüter (eds.). 2009. *One language, two grammars? differences between British and American English* (Studies in English Language). Cambridge, UK ; New York: Cambridge University Press.
Roland, Douglas & Mary Hare. 2012. Computational and Corpus Models of Human Sentence Comprehension. In Ken McRae, Marc Joanisse & Michael Spivey (eds.), *The Cambridge Handbook of Psycholinguistics*, 390–405. New York, NY: Cambridge University Press.
Rooy, Bertus Van & Lieze Terblanche. 2006. A corpus-based analysis of Involved aspects of student writing. *Language Matters* 37(2). 160–182.
Schmidtke, Daniel & Victor Kuperman. 2017. Mass counts in World Englishes: A corpus linguistic study of noun countability in non-native varieties of English. *Corpus Linguistics and Linguistic Theory* 13(1). doi:10.1515/cllt-2015-0047. https://www.degruyter.com/view/j/cllt.ahead-of-print/cllt-2015-0047/cllt-2015-0047.xml (10 October, 2018).
Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford University Press.
Sinclair, John. 2000. Lexical grammar. *Naujoji Metodologija* 24. 191–203.
Sinclair, John McH. 1998. The Lexical Item. In Edda Weigand (ed.), *Current Issues in Linguistic Theory*, vol. 171, 1. Amsterdam: John Benjamins Publishing Company. doi:10.1075/cilt.171.02sin. https://benjamins.com/catalog/cilt.171.02sin (4 October, 2018).
Stefanowitsch, Anatol & Stefan Th. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2). 209–243. doi:10.1075/ijcl.8.2.03ste.
Stubbs, Michael. 2001. *Words and phrases: corpus studies of lexical semantics*. Oxford [England] ; Malden, MA: Blackwell Publishers.
Szmrecsanyi, Benedikt. 2017. Variationist sociolinguistics and corpus-based variationist linguistics: overlap and cross-pollination potential. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 62(4). 1–17. doi:10.1017/cnj.2017.34.
Szmrecsanyi, Benedikt & Lieselotte Anderwald. 2017. Corpus-based approaches to dialect study. In Charles Boberg, John A. Nerbonne & Dominic James Landon Watt (eds.), *The handbook of Dialectology* (Blackwell Handbooks in Linguistics), 300–313. First edition. Hoboken, NJ: John Wiley & Sons, Inc.
Szmrecsanyi, Benedikt & Nuria Hernández. 2007. *Manual of Information to accompany the Freiburg Corpus of English Dialects Sampler ("FRED-S")*. Freiburg: University of Freiburg. URN: \textbackslashtexttturn:nbn:de:bsz:25-opus-28598, URL: http://www.freidok.uni-freiburg.de/volltexte/2859/.
Tagliamonte, Sali & Harald Baayen. 2012. Models, forests and trees of York English: "Was/were" variation as a case study for statistical practice. *Language Variation and Change* 24. 135–178.
Tagliamonte, Sali & Jennifer Smith. 2002. "Either it isn't or it's not": neg/aux contraction in British dialects. *English World Wide* 23(2). 251–281. doi:10.1075/eww.23.2.05tag.
Tamaredo, Iván. 2018. Pronoun omission in high-contact varieties of English: Complexity versus efficiency. *English World-Wide* 39(1). 85–110. doi:10.1075/eww.00004.tam.
Tognini-Bonelli, Elena. 2001. *Corpus linguistics at work*. Amsterdam: Benjamins.
Trudgill, Peter. 1990. *The dialects of England*. Cambridge, Mass., USA: B. Blackwell.
Tsujimura, Natsuko (ed.). 1999. *The handbook of Japanese linguistics* (Blackwell Handbooks in Linguistics). Malden, Mass: Blackwell Publishers.
Weiner, Judith & William Labov. 1983. Constraints on the agentless passive. *Journal of Linguistics* 19. 29–58.
Wetzels, W. Leo, Sergio Menuzzi & João Costa. 2016. *The Handbook of Portuguese Linguistics*. http://nbn-resolving.de/urn:nbn:de:101:1-201605043603 (23 March, 2018).
Wolk, Christoph & Benedikt Szmrecsanyi. 2018. Probabilistic corpus-based dialectometry. *Journal of Linguistic Geography* 6(1). 56–75. doi:10.1017/jlg.2018.6.
Xiao, Richard. 2015. Collocation. In Douglas Biber & Randi Reppen (eds.), *The Cambridge Handbook of English Corpus Linguistics*, 106–124. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139764377.007.

https://www.cambridge.org/core/product/identifier/9781139764377%23CN-bp-6/type/book_part (1 October, 2018).
Yáñez-Bouza, Nuria. 2011. ARCHER past and present (1990–2010). *ICAME Journal* 35. 205–236.