1

A HISTORICAL OVERVIEW OF USING CORPORA IN ENGLISH LANGUAGE TEACHING

Jiajin Xu

1.1 Introduction

Dating back to the late 1940s and the early 1950s, the Italian Jesuit priest Roberto Busa started to work on his *Index Thomisticus*, a machine-generated concordance of the writings of Saint Thomas Aquinas, cataloging 11 million Latin words with technical support from IBM (Busa, 1950, 1951; Winter, 1999). Another computerized biblical scholarship project was the concordance of the English Bible, namely, *Nelson's Complete Concordance of the Revised Standard Version Bible* (Ellison, 1957). Such pioneering work in the "digital humanities" was soon matched in linguistics by the ground-breaking Brown Corpus (Francis & Kučera, 1964; Kučera & Francis, 1967). This carefully sampled computer corpus of American written English has become the foundation of modern corpus linguistics and has foreseen a fast-growing area in linguistics over the last few decades. But the impetus for compiling text collections, corpora to support research and teaching materials can be traced back to the early 19th century. This chapter gives a historical overview of using corpora for English language teaching.

The provenance of corpus linguistics goes back a long way in history if we do not restrict it exclusively to the use of texts in electronic form (Johansson, 2011; Stubbs, 2018). In defining corpus linguistics, we share Fries' (1940, ix) argument that "[o]ne cannot produce a book dealing with language without being indebted to many who have earlier struggled with the problems and made great advances". Studies that adopt the representative sampling of authentic language data and make statistical claims of language will all be broadly regarded as examples of corpus research.

This chapter starts by sketching the history of corpus use in language-related projects, and then reviews the compilation of early English frequency lists in Section 1.2. Section 1.3 illustrates the application of corpora to the development of reference books, with special reference to pedagogical grammars and dictionaries. Section 1.4 concerns the preparation of course materials and methodological approaches using corpora. The final section outlines future avenues for corpus-informed English language teaching.

1.2 The compilation of English frequency lists

This section discusses the compilation of lexical frequency lists in the early 19th century to improve spelling skills (Section 1.2.1) and moves on to Thorndike's works since the 1920s

DOI: 10.4324/9781003002901-3

providing the impetus for idiom lists, syntax lists, and semantic frequency lists in light of corpus representativeness and range statistics (Section 1.2.2).

1.2.1 Early lexical frequency lists

In the field of English language teaching, the utilization of corpus-based quantitative methods relying on a large body of natural texts dates back to the early 19th century. Around 1820, John Freeman compiled an English frequency list based on the corpus of cca. 20,000 words to teach adults to read.[1] In 1838, Pitman (1843) developed two lists (one alphabetical and the other numerical) of frequently used words based on 10,000 words taken from 20 books written to train stenographers (shorthand writers). About half a century later, in 1897, a large-scale replica project of this stenographer-oriented frequency list, *Häufigkeitswörterbuch der deutschen Sprache*, or "A Frequency Dictionary of the German Language", was completed by Fredrick Kaeding (1897).

Since the 1910s, a large number of pre-electronic corpus projects have emerged with considerable momentum. Such early corpus research, mainly based in the United States and the Far East, was primarily motivated to facilitate language teaching.

Among the first projects of this kind were Ayres' (1913, 1915, 1920), who compiled a corpus of 2,000 personal and business letters from 12 sources, amounting to 110,160 running words. In the project reports, a Zipfian distribution (Ayres, 1913) of the most frequent words in correspondence was plotted. The most frequent word "I" implies that correspondence is a more colloquial genre than the most frequent word "the" in written or balanced corpora.

Ayres (1913, p. 10) concludes the paper by saying that,

> [t]his seems to be good evidence that a useful spelling list cannot be compiled by sitting at the desk and deciding which words people ought to know how to spell. What we must know is rather which are the words that ordinary people need to know how to spell.

This comment highlighted the striking difference between the 414 spelling words required by the National Education Association (NEA) of the United States at the time, and the actual vocabulary used by common people. Seventy percent of the NEA words "did not occur at all" (Ayres, 1913, p. 10) in the letters analyzed.

However, Ayres' quantitatively driven project was by no means the only one. Cook and O'Shea (1914, pp. 226–227) compared vocabulary found in a total of 200,000 words from the family correspondence of the 13 adults with three popular expert-compiled English spellers, and only 70 per cent of the speller words appeared in the letters. According to Cook and O'Shea (1914), the spellers did not place emphasis on what was most needed by "common people".

Similar cases include Jones' (1915) investigation and counting of 15,000,000 words of texts produced by 1,050 students from second to eighth grade and yielded 4,532 different words, i.e. word types in present-day corpus terms. Another characteristic of Jones' project is the grading of spelling vocabularies across grade levels.

The research purpose of the projects at this phase was primarily to address the problem of spelling, which was considered a central part of literacy education at school. During the next phase, we see more quantitative projects focusing on the language learning of both pupils and adults.

1.2.2 1920s–1940s: Thorndike (1921) and others

A widespread "Vocabulary Control Movement" (Cowie, 1999; Hornby, 1953; Howatt, 1984) emerged in the 1920s–1930s with more innovative curriculum goals and pedagogical practices, such as the teaching of reading and writing, syllabus design, materials development, and assessment, in addition to the teaching of spelling. The vocabulary limitation enterprise fell into two overarching approaches: one subjective and the other objective. The subjective approach was preferred by some British ELT scholars in the UK (Charles Ogden and I. A. Richards), Japan (Harold Palmer and A. S. Hornby), India, and Canada (Michael West). They adopted an intuitive approach, also known as an "armchair" approach, to the so-called BASIC Vocabulary (Ogden, 1930), standing for British American Scientific International Commercial, and the minimum adequate vocabulary (Swenson & West, 1934; West, 1931, 1934) or General Service List (Faucett *et al*., 1936; West, 1953).

American scholars, on the other hand, mainly used the objective approach to obtain minimum adequate vocabularies through quantitative methods. Thorndike's (1921) *Word Book* of 10,000 words has been regarded as a pioneering, quantitatively motivated English word list based on a large collection of authentic texts. Thorndike's work served as a key impetus for more frequency lists, not only word lists, but also idiom lists, syntax lists, and semantic frequency lists. The quantitative studies in this period outperformed those before the 1920s in terms of language varieties, aspects of language (i.e. lexis, idioms, and syntax), and improvements in methodology.

The first edition of Thorndike's *Word Book* was significantly extended from 10,000 to 20,000 words (Thorndike, 1931) and further to 30,000 words (Thorndike & Lorge, 1944) as the counts were updated and additional texts were included. Education scholars in the United States, inspired by Thorndike's work on the English language, compiled a number of frequency lists of other languages, such as French (Henmon, 1924; Vander Beke, 1929), Spanish (Buchanan, 1927), German (Morgan, 1928), and Brazilian Portuguese (Brown *et al*., 1945). This shift from mother tongue to foreign language teaching extended beyond word counts to idiom counts in French (Cheydleur, 1929), Spanish (Keniston, 1929), German (Hauch, 1929), and Brazilian Portuguese (Brown & Shane, 1951), as well as syntax counts in Spanish (Keniston, 1937) and French (Clark & Poston, 1943). Such developments in counting phraseologies and grammatical categories, however, were not reflected in the frequency counts of English.

Semantic frequency lists of English were compiled by Lorge (1937, 1949), drawing on the senses laid down in *The Oxford English Dictionary*. Lorge's semantic frequency counts were later incorporated by Faucett *et al*. (1936) in their *Interim Report on Vocabulary Selection*, and eventually published by the frequently cited West (1953) as *A General Service List of English Words: With Semantic Frequencies and a Supplementary Word-List for the Writing of Popular Science and Technology*. Moreover, the GSL (General Service List) words were set as the first two default base lists of the vocabulary tools Range (Nation, 2005) and AntWordProfiler (Anthony, 2021) developed decades later.

Almost all studies in this period took full account of text material representativeness. Keniston's Spanish corpus (1929) is a case in point as he included texts from genres such as drama, fiction, miscellaneous prose, newspapers and periodicals, and technical prose. At the same time, the sampling frame for these genres is immediately reminiscent of the widely known Brown Corpus genre categorization (Francis & Kučera, 1964), namely, press, general prose, learned (i.e. academic) writing, and fiction. In Keniston's corpus, regional varieties of Spanish (e.g. Castilian, Peninsular, and Latin American Spanish) were also considered when collecting

texts. Interestingly and surprisingly, Fries (1940) stressed, in great detail, the aspects such as authenticity, demographic representativeness, scientific sampling, and the diversity of topics/situations of corpora without adopting these contemporary corpus linguistics terms.

Another consistently followed principle in text selection and counting is range. In the introductory documentation of most studies, terms like range, distribution, "widely used", "units", and "sources" were used to illustrate this range principle alongside the frequency principle. This methodological principle was implemented since the work of Thorndike (1921) and was adhered to in almost all other studies during this period. Hence, the word/idiom/syntax counts were assigned relative frequencies as well as range statistics across different text units or sources. Furthermore, Fries (1940) called our attention to the historical differences, regional differences, literary and colloquial differences, and social and class differences in English. The differences or variational patterns of grammar points were represented by raw and relative frequencies in the 41 tables of Fries' grammar. Fries (1940) saw his book as "a study of the real grammar of [p]resent-day English [which] has never been used in the schools" (p. 285) and advised that "[w]e must agree to stimulate among our pupil[s'] observation of actual usage" (p. 291).

Regarding the aspects of language, the examination of idioms and syntax steered frequency studies beyond the word level. The idioms in question refer to both conventionalized expressions (e.g. *part and parcel*), whose meaning cannot be inferred from the component words, and lexical phrases (e.g. *pick up*). The Spanish, French, and German idiom lists were all published in 1929, which apparently predates Palmer's (1933) book-length treatment of collocations in English. Moreover, the idiom lists were based on a large quantity of natural texts and were statistically tabulated; however, Palmer's *Second Interim Report on English Collocations* was a mere list of phrases without any reference to naturally occurring texts or quantitative information. The phrasal counts of Spanish and other languages involved such constructions as compound conjunctions, compound prepositions, and verbs requiring a preposition before a complement, which naturally progressed to the quantitative description of grammatical constructions. Keniston's (1937) *Spanish Syntax List*, and Clark and Poston's (1943) *French Syntax List* followed the same range and frequency principles to quantify the full array of grammatical categories in the two languages. Besides the syntax lists, Stormzand and O'Shea (1924) took a contrastive approach to diagnosing the "excess or deficiency" (Stormzand & O'Shea, 1924, p. 48) (i.e. overuse or underuse in corpus research terms) of certain grammatical categories between adults and school children or university students. Development across grade levels was tallied and compared to gauge the progress or decline in learner performance.

Now, in the current cloud-computing age, online corpora and frequency lists are more easily available. Among these are BNC (British National Corpus) frequency lists (Leech *et al.*, 2001) and the COCA (Corpus of Contemporary American English)-based frequency dictionary of American English (Davis & Gardner, 2010). They have updated similar, previously published American English word books such as those in the 1920s–1940s and rendered them significantly more stable and reliable resources given the representativeness and size of the corpora (see Coxhead, this volume; Szudarski, this volume).

1.3 Corpora for the development of reference books

This section provides the scholarly context for the writing of pedagogical grammars, covering both systemic and comprehensive pedagogical grammars as well as some smaller and more specialized grammars (Section 1.3.1). Section 1.3.2 outlines the development of corpus-informed dictionary compilation.

1.3.1 The writing of pedagogical grammars

Unlike the numerous frequency lists, corpus-based pedagogical grammars are much less prominent during this time; they are, however, by no means insignificant. Fries' (1940) pre-electronic corpus-based English grammar is a much-neglected work. Fries, as a structuralist applied linguist, produced his *American English Grammar* on the basis of a representative corpus. The grammar book was a key reference for his seminal work titled *Teaching and Learning English as a Foreign Language* (Fries, 1945). Additionally, Fries (1940) plotted the diachronic grammatical change of English, for instance, the co-occurrence of first-, second-, and third-person pronouns with *shall* and *will* use, from 1560 to 1920. Fries (1940) also made comparative tabulations of verb prepositional/particle collocations across standard and vernacular English varieties. A decade later, Fries (1952) went further to record and transcribe the conversations of speakers of standard English in the North Central United States of cca. 250,000 words, based on which he wrote a grammar many years prior to the *Survey of English Usage* project and Quirk *et al*.'s (1972, 1985) grammars.

The Quirk-led *Survey of English Usage* and its influential Longman grammar series (1972, 1973a, 1973b, 1985, 1990) have been one of the most influential pedagogical grammar projects in the latter half of the 20th century. A clearly descriptive approach was adopted to develop the grammar. However, grammar books such as that of Quirk *et al*. (1985) do not incorporate much explicit corpus information. For instance, probabilistic information is only occasionally provided for the grammatical categories described, or with reference to the so-called The Quirk Corpus.[2]

Collins COBUILD English Grammar (Sinclair, 1990) is a systemic functional linguistics-oriented pedagogical grammar informed by corpus evidence. Typical grammatical patterns of transitivity, modality, cohesion, etc., and all the example sentences were chosen from the Birmingham Collection of English texts.

Biber *et al*. (1999) described grammatical categories and also discussed them in quantitative terms as presented in the Longman Grammar of Spoken and Written English Corpus. The other novelty in Biber *et al*.'s (1999) work was the presentation of corpus-informed register variation patterns across the conversation, news, fiction, and academic discourse in both British and American English.

The three pedagogical grammar series represented by their core grammar books, that is, Quirk *et al*. (1985), Sinclair (1990), and Biber *et al*. (1999), all have their associated concise edition or classroom edition under such names as "student grammar", "student's grammar", "basic grammar", "student grammar workbook", and "concise grammar", in order to suit classroom learning and self-study scenarios.

Apart from the big three, some smaller and more specialized corpus-based grammars also figure prominently in the ELT literature. Thornbury's (2004) *Natural Grammar* is a grammar of 100-and-something grammatical or abstract words, such as *the, do, in, much,* and *thing*. The usage of the words is presented in the form of colligations/grammatical patterns, collocation/set phrases, and example sentences/concordance lines. Conrad and Biber (2009) illustrate how register variation across speech and writing can be taught with explicit grammar patterns and situationalized activities. The typical structure of the grammar activities in the book is noticing in context, discourse-based analysis, and writing- or conversation-focused practice. McCarthy *et al*. (2009) opened up an important avenue for more grammar in the field of English for specific purposes. This *Cambridge Business Corpus-based ESP* grammar organizes major grammar points as per discourse functions or activities in business English communication. For instance, how to use the passive in business correspondence, and how to use

conditionals in business negotiation, etc. Common to the three grammar books is that they are communicatively focused and organize grammatical points according to their functions in authentic discourse.

1.3.2 The development of dictionaries

Parallel to the compilation of frequency lists and pedagogical grammar is a corpus-based approach to dictionary writing. West and Endicott's (1935) *The New Method English Dictionary* is regarded as the earliest learner's English dictionary (Cowie, 1999). The essential idea of the new method is its 1,455 most common or important defining vocabulary items (a.k.a. definition vocabulary) based on "reading counts" of language materials (West, 1935, p. 5). The New Method was also the name for a series of Longman English coursebooks and readers in which graded frequency lists were adopted to control the reading difficulty of the passages. Another equally ground-breaking learner's dictionary, *The Thorndike-Century Junior Dictionary* (Thorndike, 1935) somehow escaped lexicographical scholars' attention. Thorndike (1935) is even more corpus-informed than West and Endicott (1935) in terms of both macrostructure and microstructure arrangement. For example, entries listed in the Junior Dictionary were based on Thorndike's word books, namely, English frequency lists. At the end of each word entry, the frequency level was annotated numerically to each headword from the first thousandth (e.g. *be*…1) to the twentieth thousandth (e.g. *authorization*…20). The principle of word sense arrangement prioritizes common uses before rare uses and easily understandable uses before difficult uses, rather than in the sequence of their historical development.

The Collins COBUILD English Dictionary (Sinclair, 1987) (CCED) is probably the bona fide game changer of dictionary making in the 20th century. Corpus methodology is inherent in almost every bit of the dictionary. For instance, the selection of headwords is based on the frequency count of all English words in a 7.3-million-word corpus (initially called the Main Corpus, later referred to as Bank of English). The main innovation of the CCED is its phraseological description of the entry word. For example, the typical collocation of the word *brink*, namely, *on the brink of*, is in the first place embedded in the whole-sentence definition "If you are *on the brink of* something, usually something important, terrible, or exciting, you are just about to do it or experience it" (p. 173). The contextualized definition is itself a condensed piece of learning material. At the end of the entry, the colligational pattern is summarized as "N-SING: usu. on/to/from the N of n". The three prepositions separated by slashes are ordered according to their probability of occurrence in the corpus. Two example sentences in the same dictionary entry, namely, "Their economy is teetering *on the brink of* collapse" and "Failure to communicate had brought the two nations *to the brink of* war", were taken from the Birmingham Corpus to illustrate the characteristic uses of *on the brink of* and *to the brink of*. The co-occurrence of *brink* with *collapse* and *brink* with *war* implies the negative semantic prosody of the entry word. The extended-unit-of-meaning model, that is, the phraseological framework, has been systematically implemented in the CCED.

Recent phraseology-informed learner's English dictionaries can also be found in the EAP and ESP fields. The Louvain EAP dictionary (LEAD) (Granger & Paquot, 2015) is a web-based EAP dictionary with a special focus on collocations and recurrent phrases based on the academic component of the British National Corpus, which develops learners' awareness of discipline-specific phraseologies. The dictionary content is customizable to suit the learner's L1 background according to the information gathered from multinational learner English corpora. Discourse functions, such as defining and exemplification, are also available as starting points for dictionary lookup. Another example is Xu's (2020) work who compiled

a hotel English dictionary for tourism and hotel management students, in which frequent collocational patterns serve as a key element to link words to real-life situations. For instance, *room rate*, *room service*, and *room attendant* are listed as useful phrases underneath the headword *room*. To *do one's room* is a typical colloquial expression of hotel English, as illustrated in the example sentence "When would you like me to *do your room*, sir?" A similar domain-specific corpus approach will be adopted in 17 additional ESP dictionaries.

More recently, some advances in corpus analytical technology have expedited the writing of dictionaries. Sketch Engine is a lexicographically motivated online tool that has been adopted by major publishers. The online system can sort the typical collocations of the search word according to their grammatical relations. The fine-tuned collocations help dictionary entry writers to identify the characteristic usage patterns of target entry words. Sketch Engine also has a feature called GDEX "Good Dictionary Examples", which allows users to select dictionary friendly sentences according to criteria such as sentence length and complexity, safe topics, and the presence of difficult and low-frequency words.

In summary, corpus evidence provides quantitative information to guarantee the commonness or typicality of a word and is capable of distinguishing the senses of a word in a general or specialized domain of real-life communication.

1.4 Corpus-based materials and pedagogical approaches

This section discusses the development of course materials (Section 1.4.1) as well as methodological approaches (Section 1.4.2) using corpora. The latter includes data-driven learning and the lexical approach. Finally, research on learner corpora (Section 1.4.3) is introduced.

1.4.1 The development of course materials

The ELT course materials in this discussion mainly cover core coursebooks, supplementary materials, simplified or adapted texts, and materials evaluation. What corpora can offer to materials development includes real-life language samples, vocabulary control, a phraseological approach to lexis, and grammar. *Collins COBUILD English Course (CCEC)* is a three-level series (Willis & Willis, 1988); *Touchstone* is a four-level series (targeted at The Common European Framework of Reference (CEFR): B2 – C1) (McCarthy *et al.*, 2006, 2014); *Viewpoint* is a two-level series (targeted at CEFR: A1 – B1)[3] (McCarthy *et al.*, 2012); *On Speaking Terms: Real Language for Real Life* is a two-level series (Santana-Williamson, 2010), and *Grammar and Beyond* are a four-level series (Reppen, 2012; Reppen *et al.*, 2019). They are the major English corpus-informed coursebook series currently on the market. In *On Speaking Terms*, the content is said to have been transcribed from real-life interactions. In other words, the authenticity of the language is emphasized. However, the typicality of lexis and grammar based on the quantitative analysis of corpus data is not one of the major concerns of the coursebook design, nor is explicit information on phraseology taken into account. *On Speaking Terms* is, therefore, less representative of a corpus-based English coursebook. *CCEC* and *Touchstone* adopt the corpus approach in a more systematic manner. For instance, both series rely heavily on corpus-generated frequency lists as their primary criteria for sequencing or grading lexical and grammatical content. Thus, the scope and order of vocabulary and grammar points will be in a reasonably stepwise progression in terms of linguistic complexity. The two series both use task-oriented design to engage students in communicative activities. The listening-speaking coursebook, *Touchstone*, has an "In Conversation" section in almost every unit of the book, which is an overt illustration of how frequently a linguistic item is in the corpus of

naturally occurring discourse. For example, Unit 5 of the Touchstone Level 1 student book states that "*I mean* is one of the top 15 expressions" (p. 49). Two typical usages of *I mean* are interpreted by the Cambridge English Corpus, namely, "to repeat your ideas" and "to say more about something" (McCarthy, 2004, p. 15). These two most widely used discourse functions of *I mean* are presented to students as a must-know conversation strategy under the heading "Strategy Plus". The remainder of the section is composed of a dialog completion task and a role-play of the two conversational strategies of *I mean*. Similar corpus discoveries of spoken English are systematically incorporated into the six levels of the coursebook series, designed for learners from elementary to advanced proficiency levels.

The *Grammar and Beyond* series (Reppen, 2012) were designed as a grammar coursebook, but the exercises and/or tasks involved practice in all four language skills, with an emphasis on writing. Each unit starts with a "Grammar in the Real World" section to contextualize the use of the grammar point (e.g. demonstratives or possessives) with a real-life discourse sample. All instances of the grammar point are highlighted in boldface to enable noticing. Further corpus resources are presented in the "Data from the Real World" section in the form of charts or notes. For example, a bar chart is used to show the striking quantitative difference between indefinite pronouns with *-one* and *-body* in formal and informal registers (Reppen, 2012, p. 234). Students' attention is directed to the preference of indefinite pronouns with *one* (e.g. *someone*, *anyone*, and *everyone*) for writing and formal speaking, while indefinite pronouns ending in *-body* (e.g. *somebody*, *anybody*, and *everybody*) for informal speaking. The "Avoid Common Mistakes" section is based on the analysis of a learner corpus. Frequently committed grammatical mistakes by learners are marked with strikethroughs and correct uses are shown in different font colors.

One shared feature of *CCEC*, *Touchstone* and *Grammar and Beyond* is that authentic and typical language content is woven into a carefully crafted communicative syllabus (McCarthy, 2004).

In addition to general-purpose English coursebooks, English for Academic Purposes (EAP) coursebooks have also been developed, informed by corpus-based genre studies. Swales and Feak's (2009a, 2009b) *Michigan series in English for Academic and Professional Purposes* is a case in point. The booklets in the series focus on how to write abstracts, introductions, literature reviews, methods, results, discussions, and conclusions. The books provide a clear account of how sub-genres of research papers can be well organized by the discourse conventions of academic communities across disciplines. Language foci such as tense, reporting verb use, and genre-specific discourse strategies are summarized from authentic academic texts.

In addition to coursebook materials, corpus methods can be used in materials evaluation to measure the textual difficulty of reading passages. The Coh-Metrix (McNamara *et al*., 2014), Range (Nation, 2005), AntWordProfiler (Anthony, 2021), and Kristopher Kyle's tools (Kyle, 2021) are popular tools for analyzing reading texts and gauging their lexical, grammatical, and even discoursal features. In many cases, major ELT publishers conduct an in-house text analysis before the coursebooks are printed. Teachers and materials evaluators can assess coursebooks using on-the-fly tools.

1.4.2 Data-driven learning, the Lexcial Syllabus, and the Lexical Approach

The use of real language data and frequency lists for vocabulary or language control in ELT had been practiced long before 1990, when more systematic discussions on corpus-based syllabus design and teaching methodology were underway. Tribble and Jones (1990) started to experiment with printed concordances in language classrooms. The approach was meant to facilitate learning, and the intake of vocabulary and grammar in an inductive manner, where

linguistic meaning was derived from its context, and patterns of grammatical structures were discovered. When the approach was proposed (Johns, 1991), and later called data-driven learning (DDL), printouts of concordance lines were the primary materials for grammar and vocabulary teaching. *Collins COBUILD Concordance Samplers* (e.g. Thompson, 1995) were specifically developed for this purpose. One of the most frequently cited DDL resources is Tim Johns' Kibbitzers[4] – the language teaching materials used for EAP consultation sessions between Tim Johns and international students at Birmingham University. The dozens of the Kibbitzer cases clearly demonstrate that the DDL approach can be applied to lexical, grammatical, and discoursal levels of English teaching. Meanwhile, the native English-speaking tutor (Johns himself) did not have the final say of grammatical correctness or acceptability, but the corpus evidence, especially collocational patterns, did. The tutor and student worked together to negotiate the correct or acceptable usage against corpus resources.

The direct application of corpus resources in classrooms is connected to the Lexical Syllabus (Willis, 1990) and the Lexical Approach (Lewis, 1993). The former focuses more on the scope and sequence of lexically centered language content in ELT. The latter, however, is conceived of as an English teaching methodology parallel to the grammar-translation method, the audio-lingual method, communicative language teaching, and task-based instruction (Richards & Rodgers, 2014). Both conceptions acknowledge the relation between lexis and grammar as the two ends of a continuum. Lexis or lexical units are of central importance in English learning using this approach. Formulaic sequences or lexical phrases that are units longer than a single word are mentally stored as holistic meaning units; hence, they should be produced as a whole as well in order to achieve native-like selection and fluency (Pawley & Syder, 1983). In actual teaching, learners' awareness of formulaic sequences used in real-life discourse should be raised, and bottom-up discovery learning should be encouraged. The teaching methods are, to some extent, the blend or convergence of corpus-based phraseological analyses and task-based pedagogy.

The concordance printouts of the 1990s have now been upgraded to online DDL systems, resources, and applications. For example, web-based writing aids such as ColloCaid and Writefull, Sketch Engine for Language Learning (SkELL), Just-the-Word, StringNet, LexTutor, and Crosthwaite's short private online course (SPOC) platform, to name but a few. They can provide easy-to-generate concordances, collocational patterns, and sometimes error feedback for learners. Please refer to Part IV (Data-driven learning) in this volume for more dedicated discussions on this topic.

Earlier sections mainly focus on corpus-informed English teaching. The next section will shift to English learning in light of corpus research, especially research on learner English production.

1.4.3 Research with learner and ELF corpora

Learner corpus research (LCR) (see Part III in this volume) was initiated in Europe in the late 1980s and gained momentum in the early 1990s (Granger, 2015; Granger *et al*., 2015). Granger and her team at the Université Catholique de Louvain have contributed to the development of learner corpus compilation[5] and research. Both the design of learner corpus construction and LCR have worked in the comparative paradigm whose primary foci are the difference in English production between learners and that of the so-called native speakers, as well as the difference in English performance among learners of different first language backgrounds. The comparative methodology of LCR encapsulated in Granger's (1996b, 2015) contrastive interlanguage analysis (CIA) is still the dominant approach to LCR. It is an integrative model

of comparison with the aim of diagnosing or predicting the possible first language transfer. CIA has been updated in later years to allow for possible other dimensions (namely "reference varieties") of comparison and to accommodate the English as a Lingua Franca (ELF) view of learner English productions. Features of learner English, sometimes called errors or "foreign-soundingness" (Granger, 1996b, p. 43) in lexis, collocation, grammar, and discourse-pragmatics, can be generalized from comparisons based on corpora.

Among the most cited learner corpora in English is probably the International Corpus of Learner English (ICLE) corpus. It was built as a complementary dataset to the International Corpus of English (ICE) against a big backdrop of comparison between different varieties of English, be them native or non-native (Granger, 1996a, p. 14). Version 1 of ICLE was released in 2002, totaling 2.5 million words of essays written by learners from 11 different mother tongue backgrounds, including essays by students studying in Britain and the US. ICLE 2.0 and 3.0 were made publicly available in 2009 and 2020, respectively. A sister project, the Louvain International Database of Spoken English Interlanguage (LINDSEI), was intended to conceptualize spoken interlanguage English using a similar, comparative model, and has been extended to the Multilingual Student Translation (MUST) learner translation project, and the Longitudinal Database of Learner English (LONGDALE) project.

LCR in other regions of the world has its own localized priorities for learner corpus construction and research. For example, British and Scandinavian scholars (e.g. Nesi and Gardner, 2012 and Hasselgård, 2017, respectively) explore more research avenues of student assignments of an EAP nature. US LCR scholars (e.g. Staples *et al*., 2018) tend to consider register variation as a central concern when constructing and investigating learner corpora.

The ELF perspective of so-called non-native English production (Wu & Lei, this volume) is of special interest in the broad sense of interlanguage analysis. Examples of these are the Seidlhofer's VOICE (Vienna-Oxford International Corpus of English), Mauranen's ELFA (spoken academic English as a lingua franca), and Ishikawa's (International Corpus Network of Asian Learners of English (ICNALE) projects. This line of research impels us to reconsider contentious issues, such as non-nativeness and errors.

1.5 Future directions in corpus-informed English language teaching

The review above provides the following insights for future work: 1) More learning-driven corpora are needed. 2) More user-friendly corpus analysis tools need to be developed. 3) Experimentation into the integration of corpus resources with overall teaching objectives is strongly recommended. 4) More research on sociocultural and/or cognitive mechanisms should be carried out to validate the effectiveness of corpus application in English language teaching.

First, to facilitate learning, there is a need for more bespoke corpora that match learners' ages, current language proficiencies, and even their individualized learning needs (Jablonkai, this volume). The first two factors can be addressed by adding labels or annotations to the text in the corpus. For instance, corpus builders can take advantage of the six levels of CEFR, A1 to C2, to suggest that they are suitable for basic, intermediate, or proficient learners. In addition to the overall difficulty of the texts, text length, vocabulary coverage, and difficult word percentage can be automatically computed and marked. This provides the option for learners to be exposed to more comprehensible input texts. To cater to learners' personalized learning needs, for example, students of nursing, management, or history should be able to work with a corpus on their respective subject, or to create a sub-corpus in a larger general-purpose corpus. Lastly, the emerging multimodal corpora (e.g. including video as well as text for spoken corpora) can offer rich contextual resources for language learning.

Second, we cannot make the best of corpora for language teaching without friendly corpus tools. The current off-the-shelf software (e.g. AntConc) and online query systems (e.g. English-Corpora.org, Sketch Engine) can serve well for research purposes, but the learning curve of the tools for students is still too steep. The ideal design of a student-friendly corpus tool should be as intuitive as possible and should not require additional instruction for use; rather, it should provide sufficient contextual clues around the language item in focus (Hendry and Sheepy, this volume). Its main functionalities should cover but not be limited to frequency lists and the distribution of linguistic items in context (e.g. collocates, genre distribution). The visualization of analytical results may also be an additional highlight of the tool. The next generation of corpus tools should work with cross-platform designs, which can be used on PCs, Macs, web browsers, and mobile applications. One last point is that no matter what tool is used for classroom hands-on tasks or self-study, it should be able to foster autonomous learning, which is an inherent property of DDL (Charles, this volume).

Third, the lack of integration into the overall English curriculum might be a major drawback of the corpus approach to language teaching. More dialog and collaboration with language educators, practitioners, and ELT materials developers should be encouraged in order to bridge rich language data, and diversified, as well as individualistic learning needs. Despite the fact that corpus resources and tools can offer multiple affordances to English teaching and learning and scaffold learners at various stages of their learning, it is still questionable whether the entire English curriculum can rely on corpora. We should strive to work out an optimal mode of integration with corpus resources, methodology, and language teaching.

Fourth, the cognitive and sociological developments of corpus application to English teaching would be worthwhile topics. This research aims to explore the strengths and inadequacies of this approach. Psycholinguistic and neurolinguistic methods (e.g. reaction time, eye-tracking, and event-related potentials) as well as user logs in a registration-based web corpus system can address the issues in due course. On the learning side, the investigation of learner English has been on the cognitive aspect of learner English; for instance, conceptual metaphors (e.g. Nacey, 2013) and constructions (Gilquin, 2010) will see more of such studies. Methodologically, the multifactorial analysis (Gries, 2018; Gries *et al.*, 2020) might engender a new wave of LCR, because it considers richer contextual variables of learner performance. Gries (2018) recommends the use of regression modeling and other multivariate statistics to upgrade previous monofactorial analyses.

Acknowledgments

The author would like to acknowledge the funding provided by the Beijing Municipal Social Science Foundation project (20YYB013) "The History of Corpus Linguistics" and the support of the National Research Centre for Foreign Language Education, and the National Research Centre for State Language Capacity at Beijing Foreign Studies University. The author is extremely grateful to the editors and anonymous reviewers for their helpful comments and suggestions. The author would also like to thank Dr. Xiuling Xu and Ms. Jialei Li for reading an early version of the chapter.

Notes

1 On page 170 of a letter to the editor of *The Phonotypic Journal*, Freeman's (1820) frequency lists were reprinted.
2 Learn about the corpus at https://www.ucl.ac.uk/english-usage/about/history.htm.

3 *Viewpoint* series is the advanced level for the *Touchstone* series; hence, *Touchstone* is used subsequently to refer to all six levels of the combined series.
4 Tim Johns' Kibbitzers can be found at https://lexically.net/TimJohns/Kibbitzer/timeap3.htm.
5 "Learner corpora around the world" bookmark page at https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html.

Further reading

Friginal, E. (2018). *Corpus linguistics for English teachers: Tools, online resources, and classroom activities*. Routledge. This is an ELT teacher-friendly guidebook with rich classroom activities, lesson plans, and most importantly, step-by-step tutorials of corpus tools and resources.

Granger, S., Gilquin, G., & Meunier, F. (Eds.). (2015). *The Cambridge handbook of learner corpus research*. Cambridge University Press. This is a comprehensive handbook in which a few sections deal with similar issues to the present handbook, such as LCR and second language acquisition, LCR and language teaching.

Leńko-Szymańska, A., & Boulton, A. (Eds.). (2015). *Multiple affordances of language corpora for data-driven learning*. Benjamins. This is a collection of papers that addresses the direct use of corpora in the classroom context. The applications reported concern the improvement of speaking, writing, and translating skills, lexical and grammatical knowledge, as well as English for academic competence in light of corpora.

McCarthy, M. J., McCarten, J., & Sandiford, H. (2005). *Touchstone teacher's edition 1 with audio CD*. Cambridge University Press. This teacher's book contains the full content of *Touchstone Student's Book level 1*, the rationale for how corpus methodology is implemented in the compilation of the coursebook series, and implications for classroom use.

References

Anthony, L. (2021). AntWordProfiler (Version 1.5.1) [Computer Software]. http://www.antlab.sci.waseda.ac.jp

Ayres, L. (1913). *The spelling vocabularies of personal and business letters*. Russell Sage Foundation.

Ayres, L. (1915). *A measuring scale for ability in spelling*. Russell Sage Foundation.

Ayres, L. (1920). The spelling vocabularies of personal and business letters. *The Journal of Education*, *77*(10), 261–262, 270.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman spoken and written English grammar*. Pearson.

Brown, C., & Shane, M. (1951). *Brazilian Portuguese idiom list: Selected on the basis of range and frequency of occurrence*. Vanderbilt University Press.

Brown, C., Carr, W., & Shane, M. (1945). *A graded word book of Brazilian Portuguese*. Crofts & Co., Inc.

Buchanan, M. (1927). *A graded Spanish word book*. The University of Toronto Press.

Busa, R. (1950). Complete index verborum of works of St. Thomas. *Speculum*, *XXV*(1), 424–425.

Busa, R. (1951). *Sancti Thomae Aquinatis hymnorum ritualium varia specimina concordantiarum: Primo saggio di indici di parole automaticamente composti e stampati da macchine IBM a schede perforate* (A first example of word index automatically compiled and printed by IBM punched card machines). Fratellei Bocca.

Cheydleur, F. (1929). *French idiom list: Based on a count of 1,183,000 running words*. The MacMillan Company.

Clark, R., & Poston, L. (1943). *French syntax list: A statistical study of grammatical usage in contemporary French prose on the basis of range and frequency*. H. Holt and Company.

Conrad, S., & Biber, D. (2009). *Real grammar: A corpus-based approach to English*. Pearson Education.

Cook, W., & O'Shea, M. (1914). *The child and his spelling*. The Bobbs-Merrill Company.

Cowie, A. (1999). *English dictionaries for foreign learners: A history*. Oxford University Press.

Davis, M., & Gardner, D. (2010). *A frequency dictionary of American English: Word sketches, collocates, and thematic lists*. Routledge.

Ellison, J. (1957). *Nelson's complete concordance of the revised standard version Bible*. Thomas Nelson & Sons.

Faucett, L., Palmer, H., Thorndike, E., & West, M. (1936). *Interim report on vocabulary selection*. P. S. King & Son, Ltd.
Francis, W., & Kučera, H. (1964). *Manual of information to accompany a standard corpus of present-day edited American English, for use with digital computers*. Brown University.
Fries, C. (1940). *American English Grammar: The grammatical structure of present-day American English with especial reference to social differences or class dialects*. D. Appleton-Century-Crofts.
Fries, C. (1945). *Teaching and learning English as a foreign language*. The University of Michigan Press.
Fries, C. (1952). *The structure of English: An introduction to the construction of English sentences*. Harcourt, Brace and Company.
Gilquin, G. (2010). *Corpus, cognition and causative constructions*. Benjamins. https://doi.org/10.1075/scl.39
Granger, S. (1996a). Learner English around the world. In S. Greenbaum (Ed.), *Comparing English worldwide* (pp. 13–24). Clarendon Press.
Granger, S. (1996b). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in contrast* (pp. 37–51). Lund University Press.
Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, *1*(1), 7–24. https://doi.org/10.1075/ijlcr.1.1.01gra.
Granger, S., Gilquin, G., & Meunier, F. (Eds.). (2015). *The Cambridge handbook of learner corpus research*. Cambridge University Press.
Granger, S., & Paquot, M. (2015). Electronic lexicography goes local: Design and structures of a needs-driven online academic writing aid. *Lexicographica - International Annual for Lexicography*, *31*(1), 118–141.
Gries, S. T. (2018). On over- and underuse in learner corpus research and multifactoriality in corpus linguistics more generally. *Journal of Second Language Studies*, *1*(2), 277–309. https://doi.org/10.1075/jsls.00005.gri.
Gries, S. T., Barbara, S., Liebig, J., & Deshors, S. C. (2020). There's more to alternations than the main diagonal of a 2×2 confusion matrix: Improvements of MuPDAR and other classificatory alternation studies. *ICAME Journal*, *44*(1), 69–96.
Hasselgård, H. (2017). Stating the obvious: Signals of shared knowledge in Norwegian-produced academic English. In P. de Haan, R. de Vries, & S. van Vuuren (Eds.), *Language, learners and levels: Progression and variation* (pp. 23–44). Presses Universitaires de Louvain.
Hauch, E. (1929). *German idiom list: Selected on the basis of frequency and range of occurrence*. The MacMillan Company.
Henmon, V. (1924). *A French word book based on a count of 400,000 running words*. University of Wisconsin.
Hornby, A. S. (1953). Vocabulary control—History and principles. *ELT Journal*, *VIII*(1), 15–21.
Howatt, A. (1984). *A history of English language teaching*. Oxford University Press.
Johansson, S. (2011). A multilingual outlook of corpora studies. In V. Viana, S. Zyngier, & G. Barnbrook (Eds.), *Perspectives on corpus linguistics* (pp. 115–129). Benjamins. https://doi.org/10.1075/scl.48.08joh
Johns, T. (1991). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. *ELR Journal*, *4*, 27–45.
Jones, W. (1915). *Concrete investigation of the material of English spelling*. The University of South Dakota.
Kaeding, F. (1897). *Häufigkeitswörterbuch der deutschen Sprache*. Self-published.
Keniston, H. (1929). *Spanish idiom list: Selected on the basis of range and frequency of occurrence*. MacMillan.
Keniston, H. (1937). *Spanish syntax list: A statistical study of grammatical usage in contemporary Spanish prose on the basis of range and frequency*. H. Holt and Company.
Kučera, H., & Francis, W. (1967). *Computational analysis of present day American English*. Brown University Press.
Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: Based on the British National Corpus*. Longman.
Lewis, M. (1993). *The lexical approach: The state of ELT and a way forward*. Language Teaching Publications.
Lorge, I. (1937). The English semantic count. *Teachers College Record*, *39*(1), 65–77.
Lorge, I. (1949). *The semantic count of the 570 commonest English words*. Teachers College, Columbia University.
McCarthy, M. (2004). *Touchstone: From corpus to course book*. Cambridge University Press.
McCarthy, M., McCarten, J., Clark, D., & Clark, R. (2009). *Grammar for business*. Cambridge University Press.

McCarthy, M., McCarten, J., & Sandiford, H. (2006). *Touchstone student book level 1*. Cambridge University Press.
McCarthy, M., McCarten, J., & Sandiford, H. (2012). *Viewpoint level 1 student's book*. Cambridge University Press.
McCarthy, M., McCarten, J., & Sandiford, H. (2014). *Touchstone student book level 1* (2nd ed.) Cambridge University Press.
McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.
Morgan, B. (1928). *German frequency word book: Based on Kaeding's Häufigkeitswörterbuch der deutschen Sprache*. MacMillan.
Nacey, S. (2013). *Metaphors in learner English*. Benjamins.
Nation, I. (2005). Range [Computer Software]. https://www.wgtn.ac.nz/lals/resources/paul-nations-resources
Nesi, H., & Gardner, S. (2012). *Genres across the disciplines: Student writing in higher education*. Cambridge University Press.
Ogden, C. (1930). *The basic vocabulary: A statistical analysis*. Kegan Paul, Trench, Trubner & Co, Ltd.
Palmer, H. (1933). *Second interim report on English collocations*. The Institute for Research in English Teaching, Department of Education.
Pawley, A., & Syder, F. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191–226). Longman.
Pitman, I. (1843). List of words from which grammalogues may be selected. *The Phonotypic Journal*, *2*(23), 161–163.
Quirk, R., & Greenbaum, S. (1973a). *A university grammar of English*. Longman.
Quirk, R., & Greenbaum, S. (1973b). *A concise grammar of contemporary English*. Harcourt Brace Jovanovich.
Quirk, R., & Greenbaum, S. (1990). *A student's grammar of the English language*. Longman.
Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1972). *The grammar of contemporary English*. Longman.
Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *The comprehensive grammar of the English language*. Longman.
Reppen, R. (2012). *Grammar and beyond level 1 student's book*. Cambridge University Press.
Reppen, R., Blass, L., Iannuzzi, S., Savage, A., Bunting, J. D., & Diniz, L. (2019). *Grammar and beyond essentials*. Cambridge University Press.
Richards, J. C., & Rodgers, T. S. (2014). *Approaches and methods in language teaching*. Cambridge University Press.
Santana-Williamson, E. (2010). *On speaking terms 1: Real language for real life*. Cengage Learning.
Sinclair, J. (1987). *Collins COBUILD dictionary of English language*. Collins.
Sinclair, J. (1990). *Collins COBUILD English grammar*. Collins.
Staples, S., Biber, D., & Reppen, R. (2018). Using corpus-based register analysis to explore the authenticity of high-stakes language exams: A register comparison of TOEFL iBT and disciplinary writing tasks. *The Modern Language Journal*, *102*(2), 310–332. https://doi.org/10.1111/modl.12465.
Stormzand, M., & O'Shea, M. (1924). *How much English grammar? An investigation of the frequency of usage of grammatical constructions in various types of writing together with a discussion of the teaching of grammar in the elementary and the high school*. Warwick & York, Inc.
Stubbs, M. (2018). The (very) long history of corpora, concordances, collocations and all that. In A. Čermáková & M. Mahlberg (Eds.), *The corpus linguistics discourse: In honour of Wolfgang Teubert* (pp. 9–33). Benjamins. https://doi.org/10.1075/scl.87.02stu
Swales, J., & Feak, C. (2009a). *Abstracts and the writing of abstracts*. The University of Michigan Press.
Swales, J., & Feak, C. (2009b). *Telling a research story: Writing a literature review*. The University of Michigan Press.
Swenson, E., & West, M. (1934). *On the counting of new words in textbooks for teaching foreign languages*. The University of Toronto Press.
Thompson, G. (1995). *Collins COBUILD concordance samplers: Reporting*. HarperCollins Publishers Ltd.
Thornbury, S. (2004). *Natural grammar: The keywords of English and how they work*. Oxford University Press.

Thorndike, E. (1921). *The teacher's word book*. Columbia University.
Thorndike, E. (1931). *A teacher's word book of the twenty thousand words: Found most frequently and widely in general reading for children and young people*. Columbia University.
Thorndike, E. (1935). *The Thorndike-century junior dictionary*. D. Appleton-Century-Crofts.
Thorndike, E., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. Columbia University.
Tribble, C., & Jones, G. (1990). *Concordances in the classroom*. Longman.
Vander Beke, G. (1929). *French word book*. MacMillan.
West, M. (1931). Notes, news and clippings. *The Modern Language Journal*, *15*(8), 638–647.
West, M. (1934). English as a world language. *American Speech*, *9*(3), 163–174.
West, M. (1935). *Definition vocabulary*. University of California, Davis.
West, M. (1953). *A general service list of English words: With semantic frequencies and a supplementary word-list for the writing of popular science and technology*. Longmans, Green.
West, M., & Endicott, J. (1935). *The new method English dictionary*. Longman.
Willis, D. (1990). *The lexical syllabus: A new approach to language teaching*. Collins.
Willis, J., & Willis, D. (1988). *Collins COBUILD English course: Student's book 1*. Collins.
Winter, T. (1999). Roberto Busa, S.J., and the invention of the machine-generated concordance. *The Classical Bulletin*, *75*(1), 3–20.
Xu, J. (2020). *Xin shidai zhiye yingyu jiudian yingyu cihui shouce* [New era vocational English word book: Hotel English]. Foreign Language Teaching and Research Press.