What's out there? A literature study on the typology of English corpora and their contribution to ELT[1]

Srifani Simbuka
Institut Agama Islam Negeri Manado
Universitas Pendidikan Bandung
Manado-Bandung, Indonesia
srifani.simbuka@iain-manado.ac.id

Fuad A. Hamied
Universitas Pendidikan Indonesia
Bandung, Indonesia
fuadah@upi.edu

Wachyu Sundayana
Universitas Pendidikan Indonesia
Bandung, Indonesia
swachyu@upi.edu

Deny A. Kwary
Universitas Airlangga
Surabaya, Indonesia
kwary@yahoo.com

Abstract

One of the recent challenges for ELT practitioners is to respond to innovations in technology such as the growth of web-based English corpora. These corpora are rich resources for ELT since they contain most current, empirical data on the dynamic English language. The purpose of this paper is to review the literature on the typology of English corpora by describing and exemplifying "what English corpora are out there" for ELT practitioners to explore and harness their benefits. The analysis was conducted by collecting and selecting literature containing the concepts and characteristics of corpora. The result is a synthesis of the literature outlining three main bases for determining the types of corpora. First, there is a criterion related to the design and use of corpora and refers to the 'age' of their sources. The second criterion is the scope of corpora, and the third is 'flexibility' toward changes in their surroundings. Based on the first criteria, corpora are labeled as synchronic from diachronic corpora. By the second criteria, corpora are distinguished into general versus specific ones. Static versus dynamic corpora are the types of corpora that indicated their 'flexibilities'. Based on this result, it is recommended that corpora users are fully aware of the specific characteristics of each type of corpora, and hence they could choose the ones that best suit their needs.

[1] This paper is a part of a dissertation entitled "The Islamic Religious Studies Technical Vocabulary (IRSTV): A Corpus-Based Study to Inform Materials Development", at the UPI Bandung EED Program, written by the first author and supervised by the other authors, forthcoming.

Keywords: English corpora, general/specific corpora, synchronic/diachronic corpora, static/dynamic corpora, typology of corpora

Introduction

This paper sketches the literature on the typology of corpora or the collections of (usually) massive language data, which are collected using principled methods and saved in computers or digital storing and processing media. English corpora (singular: corpus), created under the Corpus Linguistics, have contributed to English Language Teaching (ELT) mainly by providing empirical linguistic data as the basis for compiling dictionaries and word/vocabulary lists to assist learners in their ESL/EFL endeavors (Nation, 2016; Xiao & McEnery, 2012). Advances in technology have allowed corpora to harness more sophisticated and cut-edge digital medium for improving the quality of corpora. Eventually, ELT programs that make use of the findings of corpus-based research in their syllabuses and materials will also be improved.

Prior to using corpora for any pedagogical purposes, it is imperative that ELT practitioners are aware of the classification of corpora that are mostly available online. Choosing the 'right' corpus for an ELT program would be made simpler, should ELT practitioners are equipped with sufficient knowledge of the typology of corpora, their specific characteristics and some examples of each of the corpora types. The sections that follow provide a summary of this information on corpora, which is actually a fraction of the literature review of a doctorate research report written by the authors.

Corpus: a brief review

Scholars have agreed collectively to describe corpus as "a set of genuine language, either written or spoken, compiled for a specific purpose" (Flowerdew, 2012 quoting Sinclair, 1991 ; (Biber, 2015). Weisser, (2016) narrows this 'particular purpose' stating that a corpus is "… any collection of texts that have been systematically assembled in order to investigate one or more linguistic phenomena … even if it may only contain a handful of classroom transcripts, interviews, or plays." (Weisser, 2016: 3). The ESRC Centre for Corpus Approaches to Social Science (CASS) (2013: 5) defined corpus as:

> "From the Latin for 'body' (plural corpora), a corpus is a body of language representative of a particular variety of language or genre which is collected and stored in electronic form for analysis using concordance software."

An important feature of a corpus (plural: corpora) is that the language that built a corpus or corpora should be language or languages that occur naturally and collected under a principled linguistic design (Abudukeremu, 2010; Biber, 2015; Flowerdew, 2012). This principled data collection is the one feature of corpus/corpora that distinguishes a corpus from a database and or web/world wide web (Flowerdew, 2012).

The typology of corpora

Corpora are grouped into three categories based on different key characteristics Weisser (2016). The first characteristic of corpora that differentiates one category from another is related to the design and use (Weisser, 2016). This characteristic is also related to the 'age' of their sources. Synchronic and diachronic corpora are the two opposing types of corpora based on this first characteristic (Weisser, 2016). The second characteristic has to do with the coverage of corpora that distinguishes corpora into general versus specific ones. Finally, corpora are classified based on their 'flexibility' toward the shifts in their surrounding environments. Based on this feature, corpora are classified into static versus dynamic corpora.

Synchronic versus diachronic corpora

By Weisser's (2016) typology, synchronic corpora whose purpose of design is to portray modern-living languages are separated from diachronic ones. Diachronic corpora are therefore designed to historically describe archaic languages that have already extinct. A further difference between these two contrasting types of corpora is the nature of their data sources. Synchronic corpora are built of the data of currently used languages. On the contrary, diachronic corpora use data of historical or archaic languages, such as “Shakespeare’s English”. This type of corpora usually contains antiquated words, words with ‘old-style’ spelling or characters that are unknown and or have died out in modern-day texts.

Synchronic corpora are built of written texts, spoken texts, and texts that are the combination of these two modes. Meanwhile, the texts that built diachronic corpora are mostly written ones, due to the unavailability of recording devices that could capture the spoken variation of the archaic languages from which diachronic corpora are drawn their data.

Written versus spoken corpora

The modes of the data source of synchronic corpora further separate them into written, spoken and mixed-mode ones. Written corpora usually are larges in size and more varied in genres, while spoken corpora are in the opposite pole. The disparity of size and genres of these two types of corpora is caused by difficulty in data collection and processing of spoken texts. Written texts for written corpora do not contain supra-segmental features such as intonation, pause or hesitation, dialect markers or any extra-linguistic elements of texts that exist in spoken texts.

The Brown corpus is one of the examples of pioneering written corpora, compiled in the 1960s, is the first-ever computerized corpus. This corpus is named after the location of the development that is, Brown University (USA). The Brown corpus developed by Nelson Francis and Henry Kucˇera is published in 1964 and still operating (Flowerdew, 2012). It includes one million American English phrases (Weisser, 2016) using as its information files released in 1961. The Lancaster-Oslo-Bergenor or the LOB available at http:/clu.uni.no/icame/manuals/LOB/INDEX.HTM is another instance. The 1978 LOB is the British counterpart to Brown Corpus. Both the

Brown Corpus and the LOB are regarded significant corpora of the first generation (Flowerdew, 2012).

Some early spoken corpora include the Survey of English Usage or SEU. The size of the SEU is one million words. The British counterpart of the SEU is the London-Lund Corpus or LLC (http://clu.uni.no/icame/manuals/LONDLUND/INDEX.HTM).

Mixed corpora

Mixed corpora, as the name suggests, contain both written and spoken corpora. By creating a balance between spoken and written language, these corpora strive to be more representative of language in particular. Although the sizes of some of the mixed-corpora are originally small, they develop into hundreds of millions of words of the latest mega corpora.

This growth is made possible by the advance in data collection (Weisser, 2016). the American National Corpus (ANC), which is the American counterpart of the BNC, is one of the examples of mega corpora. The ANC is available, although only partially free, from http://americannationalcorpus.org/. The ANC has reached 22 million words in size and continues to evolve. Another example is the Corpus of Contemporary American English (COCA) developed by Mark Davies at the Birmingham Young University/BYU. This corpus of American English is available from http://corpus.byu.edu/coca/. Its data source is a balanced collection of spoken materials from various genres such as fiction, popular magazines, newspapers, and academic texts. It includes over 450 million words of text, most of which can be freely accessed through the internet interface. Because of its increasing size, COCA has worked as a monitor corpus. The Corpus of Global Web-based English (GloWbE) http:/corpus.byu.edu/glowbe/is regarded a mega corpus as it holds a 1.9 billion word collection. This collection is web pages of English from 20 countries, both native and non-native variations.

As stated above, diachronic corpora are corpora devoted to the research of archaic/historical texts. Some of the examples of this category are the Helsinki corpus and the Corpus of Historical American English (COHA). The Corpus of Helsinki maintains historical materials from ca. 750–1700, plus a number of English rural dialect transcripts from the 1970s. This corpus is available from http://clu.uni.no/icame/manuals/HC/ INDEX.HTM and also through the Oxford Text Archive at http://ota.ahds.ac.uk/headers/ 1477.xml. The COHA holds texts dating from the early 19th century to early millennial (1810 to 2009). This corpus is 400 million words in size and can be found at http:/corpus.byu.edu/coha/ (Weisser, 2016). Also listed under this category are diachronic corpora based on texts from other languages such as the work of Roberto Busa (Flowerdew, 2012). In a project sponsored by a renowned computer business, IBM, Busa created a corpus of medieval philosophy texts using a concordancer.

General versus specific corpora

Weisser (Weisser, 2016) also makes a distinction of corpora based on the scope of their purposes, contrasting general from domain-specific corpora (Weisser, 2016). General corpora cover as many language varieties, mode, and text genres in order that these corpora be deemed representative of the whole language. Hence, one individual corpus of this type of corpora is used in many research addressing various objectives. Unlike general corpora, Domain-specific corpora or field-specific corpora often involve particular linguistic variations, modes or genres (Weisser, 2016). These types of corpora are usually much smaller in size compared to the general ones. Since corpus size is not a salient factor for judging the representativeness of this type of corpora, therefore, the representativeness of field-specific corpora is defined by the research questions (Nation, 2016) or the purposes of their development. Domain-specific corpora are thus useful because they can highlight particular aspects of the examined language, for instance, by pointing out differences between standard language and specific registers.

While most synchronic (written, spoken or mixed) corpora are constructed for general research interests, there are some of these corpora that fall into the sub-type of specific corpora. These are academic corpora, learners' corpora, and pragmatically annotated corpora. Academic corpora are those containing language exclusively produced in academic contexts, i.e. English for Academic Purposes or EAP (Nation, 2016; Weisser, 2016). The British Academic Written English (BAWE) is one of the examples of academic corpus developed using assignments written by high achiever-native speakers of English students from three UK universities. Learners' corpora differ from academic corpora in that the data source of this type of corpora is, as the name suggests, learners at various levels and stages of language acquisition and not an expert in the field. The texts used as data source of learners' corpora are often limited to the works of L2 learners or non-native speakers of a language (Weisser, 2016). Learners' corpora are often the source of many interesting topics in pedagogy-oriented corpus research as they provide rich data on the pattern of L2 indicating learners' progress towards the L2 learning/acquisition process (Flowerdew, 2012; O'Keeffe & McCarthy, 2010). The first corpus of a special variety of a language was the Jiao Da English for Science and Technology (JDEST) corpus, a work of Yang Huizhong in Shanghai (O'Keeffe & McCarthy, 2010; Vivanet, 2012). Finally, pragmatically annotated corpora are deemed special ones due to their rather 'unusual' pragmatic tagging for speech acts or other pragmatically relevant information (Weisser, 2016) instead of 'the normal' form/grammatical tagging for parts of speech (POS).

Static versus dynamic corpora

With regard to the flexibility of corpora to grow in size, Weisser (2016) distinguishes between snapshot versus monitor corpora. The main differences between these two sub-types of corpora lie on the time frame used in the data collection and on the size. Monitor corpora are constructed according to a specific sampling frame and gradually grow in to include more texts over time at a given point in time (McEnery & Hardie, 2012). It is the purpose of monitor corpora to reflect the language they portray. Hence, the sizes monitor corpora tend to increase over time, while the size of snapshot corpora are fixed. The changes in monitor corpora are expected since they are designed to keep mirroring, thus, monitor, the dynamic change of language they represent. Until recently, the COCA and the Bank of English (BOE) are the only two genuine monitor corpora in existence (Weisser, 2016).

Snapshot corpora or 'sample corpora' (McEnery & Hardie, 2012), was designed to capture the language they represent at a specific time span within the language's own history. This type of sub-corpora pictured a specific phenomenon that happened in a language in a particular 'event', such as, a change in a language due to a shift in its social, political or economic surroundings.

Recent corpora

Current development in the theory of Corpus Linguistics and the advances in technology that supported corpora constructions have allowed new corpora to flourish. As a result, single-typed corpora exemplified in the elaboration above have also become lesser in number, given that new sources, techniques, and tools for data source retrieval and processing are available. Recently built corpora, therefore, establish themselves as 'multi-typed' corpora, having many cross-over characteristics that blurred the boundary of the conventional typology of corpora.

Some of the current trends in corpora development are the growing attention to creating specific corpora addressing topics related to field-specific area. At the same time, these corpora are snapshot synchronic ones. And yet, the written mode of data was still dominant over the spoken one. As a result, the literature recorded more corpora having a combined type of synchronic-specific-written characteristics. Corpora of this hybrid nature are mostly created as the basis of technical word lists creation.

Some of the instances of this multi-typed corpora are the ESP corpus/ESPC and EFL corpus/EFLC (Shabani & Tazik, 2014), the corpus of plumbing (Coxhead & Demecheleer, 2018). the corpus of Islamic Academic Research Article/IARA (Ibrahim, Shah, & Abudukeremu, 2019), and the Corpus of Islamic Religious Study Textbooks in Indonesian State Institute for Islamic Studies/ CIRST-ISIIS (Simbuka, Hamied, Sundayana, & Kwary, 2019). The main thread that connects these examples of multi-typed corpora is the nature of their purpose of development, that is, to be used as the springboard for creating domain-specific word lists of technical vocabulary. The end products of these corpora are word lists beneficial for ELT.

Conclusion

The typology of corpora suggested by Weisser (2016) has outlined some basic characteristics of most existing corpora. Some of these characters may be combined in certain corpora, making them multi-typed corpora. Most of the 'modern' corpora can be attributed as 'synchronic-mixed-general-static ones, for instance, the BNC and the ANC. Meanwhile, others can be labeled as 'synchronic-written (or spoken)-specific-static' corpora such as the BAWE. Based on this, it can be underlined that most recent corpora fall into a combined typology of synchronic-static corpora. The sub-typology of general versus specific attributes are then used as the one criteria that differentiate one corpus from the others.

References

Abudukeremu, M. (2010). *A Corpus-Based Lexical Study of the Frequency, Coverage, and Distribution of Academic Vocabulary in Islamic Academic Research Articles*. International Islamic University Malaysia.

Biber, D. (2015). *Oxford Handbooks Online Corpus-Based and Corpus-Driven Analyses of Language Variation and Use*.

https://doi.org/10.1093/oxfordhb/9780199677078.013.0008

Coxhead, A., & Demecheleer, M. (2018). English for Specific Purposes Investigating the technical vocabulary of Plumbing. *English for Specific Purposes*, *51*, 84–97. https://doi.org/10.1016/j.esp.2018.03.006

Flowerdew, L. (2012). *Corpora and Language Education*. Palgrave Macmillan.

Ibrahim, E. H. E., Shah, M. I. A., & Abudukeremu, M. (2019). A Corpus-Based Lexical Study of the Frequency, Coverage, and Distribution of Academic Vocabulary in Islamic Academic Research Articles. *The Journal of Social Sciences Research*, (SPI 2), 570–577. https://doi.org/10.32861/jssr.spi2.570.577

McEnery, T., & Hardie, A. (2012). Corpus linguistics : method, theory, and practice. *Cambridge Textbooks in Linguistics*, xv, 294 p. https://doi.org/10.1017/CBO9781107415324.004

Nation, I. S. P. (2016). *Making and Using Word Lists for Language Learning and Testing*. (I. S. P. Nation, Ed.) (e-Book). John Benjamins. https://doi.org/10.1075/z.208

O'Keeffe, A., & McCarthy, M. (2010). *The Routledge Handbook of Corpus Linguistics*. (A. O'Keeffe & M. McCarthy, Eds.) (1st ed.). New York: Routledge.

Shabani, M. B., & Tazik, K. (2014). Coxhead ' s AWL across ESP and Asian EFL Journal Research Articles (RAs): A Corpus-Based Lexical Study. *Procedia - Social and Behavioral Sciences*, *98*, 1722–1728. https://doi.org/10.1016/j.sbspro.2014.03.599

Simbuka, S., Hamied, F. A., Sundayana, W., & Kwary, D. A. (2019). A Corpus-Based Study on the Technical Vocabulary of Islamic Religious Studies. *TEFLIN Journal*, *30*(1), 47–71. https://doi.org/http://dx.doi.org/10.15639/teflinjournal.v30i1/47-71

The ESRC Centre for Corpus Approaches to Social Science (CASS). (2013). *Corpus : Some key terms* (No. 1). Lancaster.

Vivanet, G. (2012). Book Review: The Routledge Handbook of Corpus Linguistics. *Humana Mente Journal of Philosophical Studies*, *23*, 183–188.

Weisser, M. (2016). *Practical Corpus Linguistics: An Introduction to Corpus-Based Language Analysis* (1st ed.). Malden: Wiley & Blackwell. https://doi.org/10.1002/9781119180180

Xiao, Z., & McEnery, T. (2012). *Corpora and language education*. *English for Specific Purposes* (Vol. 23). https://doi.org/10.1016/j.esp.2003.12.001

Abudukeremu, M. (2010). *A Corpus-Based Lexical Study of the Frequency, Coverage, and Distribution of Academic Vocabulary in Islamic Academic Research Articles*. International Islamic University Malaysia.

Biber, D. (2015). *Oxford Handbooks Online Corpus-Based and Corpus-Driven Analyses of Language Variation and Use*. https://doi.org/10.1093/oxfordhb/9780199677078.013.0008

Coxhead, A., & Demecheleer, M. (2018). English for Specific Purposes Investigating the technical vocabulary of Plumbing. *English for Specific Purposes*, *51*, 84–97. https://doi.org/10.1016/j.esp.2018.03.006

Flowerdew, L. (2012). *Corpora and Language Education*. Palgrave Macmillan.

Ibrahim, E. H. E., Shah, M. I. A., & Abudukeremu, M. (2019). A Corpus-Based Lexical Study of the Frequency, Coverage, and Distribution of Academic Vocabulary in Islamic Academic Research Articles. *The Journal of Social Sciences Research*, (SPI 2), 570–577. https://doi.org/10.32861/jssr.spi2.570.577

McEnery, T., & Hardie, A. (2012). Corpus linguistics : method, theory, and practice. *Cambridge Textbooks in Linguistics*, xv, 294 p. https://doi.org/10.1017/CBO9781107415324.004

Nation, I. S. P. (2016). *Making and Using Word Lists for Language Learning and Testing*. (I. S. P. Nation, Ed.) (e-Book). John Benjamins. https://doi.org/10.1075/z.208

O'Keeffe, A., & McCarthy, M. (2010). *The Routledge Handbook of Corpus Linguistics*. (A. O'Keeffe & M. McCarthy, Eds.) (1st ed.). New York: Routledge.

Shabani, M. B., & Tazik, K. (2014). Coxhead ' s AWL across ESP and Asian EFL Journal Research Articles (RAs): A Corpus-Based Lexical Study. *Procedia - Social and Behavioral Sciences*, *98*, 1722–1728. https://doi.org/10.1016/j.sbspro.2014.03.599

Simbuka, S., Hamied, F. A., Sundayana, W., & Kwary, D. A. (2019). A Corpus-Based Study on the Technical Vocabulary of Islamic Religious Studies. *TEFLIN Journal*, *30*(1), 47–71. https://doi.org/http://dx.doi.org/10.15639/teflinjournal.v30i1/47-71

The ESRC Centre for Corpus Approaches to Social Science (CASS). (2013). *Corpus : Some key terms* (No. 1). Lancaster.

Vivanet, G. (2012). Book Review: The Routledge Handbook of Corpus Linguistics. *Humana Mente Journal of Philosophical Studies*, 23, 183–188.

Weisser, M. (2016). *Practical Corpus Linguistics: An Introduction to Corpus-Based Language Analysis* (1st ed.). Malden: Wiley & Blackwell. https://doi.org/10.1002/9781119180180

Xiao, Z., & McEnery, T. (2012). *Corpora and language education*. *English for Specific Purposes* (Vol. 23). https://doi.org/10.1016/j.esp.2003.12.001

Abudukeremu, M. (2010). *A Corpus-Based Lexical Study of the Frequency, Coverage, and Distribution of Academic Vocabulary in Islamic Academic Research Articles*. International Islamic University Malaysia.

Biber, D. (2015). *Oxford Handbooks Online Corpus-Based and Corpus-Driven Analyses of Language Variation and Use*. https://doi.org/10.1093/oxfordhb/9780199677078.013.0008

Coxhead, A., & Demecheleer, M. (2018). English for Specific Purposes Investigating the technical vocabulary of Plumbing. *English for Specific Purposes*, *51*, 84–97. https://doi.org/10.1016/j.esp.2018.03.006

Flowerdew, L. (2012). *Corpora and Language Education*. Palgrave Macmillan.

Ibrahim, E. H. E., Shah, M. I. A., & Abudukeremu, M. (2019). A Corpus-Based Lexical Study of the Frequency, Coverage, and Distribution of

Academic Vocabulary in Islamic Academic Research Articles. *The Journal of Social Sciences Research*, (SPI 2), 570–577. https://doi.org/10.32861/jssr.spi2.570.577

McEnery, T., & Hardie, A. (2012). Corpus linguistics : method, theory, and practice. *Cambridge Textbooks in Linguistics*, xv, 294 p. https://doi.org/10.1017/CBO9781107415324.004

Nation, I. S. P. (2016). *Making and Using Word Lists for Language Learning and Testing*. (I. S. P. Nation, Ed.) (e-Book). John Benjamins. https://doi.org/10.1075/z.208

O'Keeffe, A., & McCarthy, M. (2010). *The Routledge Handbook of Corpus Linguistics*. (A. O'Keeffe & M. McCarthy, Eds.) (1st ed.). New York: Routledge.

Shabani, M. B., & Tazik, K. (2014). Coxhead ' s AWL across ESP and Asian EFL Journal Research Articles (RAs): A Corpus-Based Lexical Study. *Procedia - Social and Behavioral Sciences*, 98, 1722–1728. https://doi.org/10.1016/j.sbspro.2014.03.599

Simbuka, S., Hamied, F. A., Sundayana, W., & Kwary, D. A. (2019). A Corpus-Based Study on the Technical Vocabulary of Islamic Religious Studies. *TEFLIN Journal*, *30*(1), 47–71. https://doi.org/http://dx.doi.org/10.15639/teflinjournal.v30i1/47-71

The ESRC Centre for Corpus Approaches to Social Science (CASS). (2013). *Corpus : Some key terms* (No. 1). Lancaster.

Vivanet, G. (2012). Book Review: The Routledge Handbook of Corpus Linguistics. *Humana Mente Journal of Philosophical Studies*, *23*, 183–188.

Weisser, M. (2016). *Practical Corpus Linguistics: An Introduction to Corpus-Based Language Analysis* (1st ed.). Malden: Wiley & Blackwell. https://doi.org/10.1002/9781119180180

Xiao, Z., & McEnery, T. (2012). *Corpora and language education*. *English for Specific Purposes* (Vol. 23). https://doi.org/10.1016/j.esp.2003.12.001