Corpus Linguistics in Teaching, Learning, and Research

تنفيذ اللغويات النصوصية في التدريس و التعلم و البحث

La linguistique de corpus dans l'enseignement, l'apprentissage et la recherche

Nadjouia RAOUD
Alger 2جامعة الجزائر

Introduction

Corpus linguistics (hereafter sometimes referred to as CL) is the study and analysis of machine readable data obtained from a variety of corpora, which are collections of written as well as transcribed texts from spoken language in a variety of domains. This article discusses the question of its implementation in teaching/learning as one of the fundamental educational, pedagogic, and academic requirements. It attempts to shed light on the various ways in which corpus linguistics has revolutionalised language instruction and research. The availability of various types of corpora has made it possible to access real-life use of language in an array of genres and registers, ranging from such domains as journalism, fiction, learned as well as interlanguage texts. Among the first to introduce corpus linguistics are Higgings & Johns (1984), Johns (1986), and Tribble & Jones (1990, as cited in Pérez-Paredes, 2010). The foundation of corpus linguistics was laid in the 90's with Sinclair's pioneering work (1991), followed in the 2000's by a course hosted in the Tuscan Word Centre with the title *How to Use Corpora in Language Teaching*, leading to a book edited by Sinclair under the same title (Sinclair, 2004). Since then published literature in the field of corpus linguistics abounds, for example, Aston (2001), Barlow (2002), Hunston (2002), Krieger (2003), and Sinclair (1991, 2003, 2004), Meunier (2010), but also several tutorials on how to use the software required for its application. More recent publications include Baker & Egbert (2016) on different corpus linguistic methodologies and triangulation in corpus-related research and McCarthy (2022), which provides a wealth of papers on applications, analyses, and theoretical issues pertaining to corpus linguistics.

1. Types of Corpora

Corpus linguistics is not new. As far back as the 19 th century, linguists resorted to large collections of texts for various analyses, a task which proved tiresome and time-consuming. With the advent of technology and globalisation,

Reçu le 09-07-2021. pub-num 28-02-2021. Pub-Asjp. le 08-03-2022
https://aleph-alger2.edinum.org/5936

access to countless representative instances of authentic written and spoken language formerly inaccessible has been made customary through computers, and several types of corpora have emerged (cf. University of Essex, *Corpora,* for a detailed account of types of corpora).

1.1. First generation corpora

The first technology-based corpus, known as the one-million-word Brown Corpus and created in 1961 by Henry Kučera and W. Nelson Francis at Brown University, Rhode Island, USA, contained 500 texts ranging from journalistic reports and reviews to humour and learned language through literary science and mystery fiction texts each consisting of about 2000 words. Other corpora have folllowed since: the LOB, (Lancaster-Oslo-Bergen) Corpus (British English), the Kolhapur Corpus (Indian English), and the LLC, the London-Lund Corpus of Spoken British English. The latter, directed by J. Svartvik, which was the first computer readable corpus of spoken language, consists of 100 spoken texts -recorded since the 1950s- that contain about 5,000 words each. These texts, orthographically transcribed, are classified into different categories : spontaneous conversations and commentary, as well as spontaneous and prepared oral texts. Despite their usefulness, these first-generation corpora were however soon challenged by larger and more up to date corpora. The COBUILD (Collins Birmingham University International Language Database) corpus, initially led by Professor John Sinclair, created in the 1980s with the aim of compiling dictionaries, allowed continous and daily access of approximately 20 million words to the Collins Cobuild English Language Dictionary makers. In the eighties, the author of this article witnessed unparalleled enthusiasm around the compilation of the Collins English Dictionary published in 1983. The BoE, (Bank of English), launched in1991, contained in 1996 some 320 million words. These two corpora (COBUILD and BoE) are known as monitor (non finite) corpora because new words are constantly being added. In 1995 another large finite corpus of written and spoken English was released, the BNC (British National Corpus), consisting of some 100 million words.

1.2 Specialized corpora

Among the specialised corpora are the historical corpora, such as the Helsinki Corpus of English Texts, a diachronic 1,5 million word corpus of English texts dating from Old, Middle and Early Modern English ; the Lampeter Corpus of Early Modern English Tracts, consisting of tracts and pamphlets published between 1640 and 1740 ; the Corpora for Special Purposes such as the ATCO (Air Traffic Control Corpus) and the Trains Spoken Dialogue Corpus, created

in the railway freight system as a conversation assistant, a number of which are recorded tasks by individuals, for example, telephone conversations, name spelling and repeating words and numbers, where people are asked to perform a particular task over the telephone, such as saying and spelling their name or repeating certain words, phrases, numbers, and letters ; and the LDC (Linguistic Data Consortium), an open consortium of universities and research laboratories, hosted by the University of Pennsylvania, which collects and distributes speech and text databases and other linguistics resources for research and development purposes. Other specialised corpora involve international or multilingual corpora, the latter used in machine translation, for example, the parallel corpora, corpora bearing several translations of the same text. The International Corpus of English, which comprises native speaker language, a British component (ICE-GB), consists of 20 corpora of different international native speaker varieties of English while the International Corpus of Learner English is produced by (learner) non-native speakers of English in the different countries. Several other corpora have emerged ; for example, the Canterbury Tales Project, which offers transcripts of the Canterbury Tales, Chaucer's Middle English prose and verse ; CSLU : The Center for Spoken Language Understanding, a multidisciplinary center for language and biomedical research ; and TELRI (Trans-Euràpean Language Resources Infrastructure), a European Commission run project for supplying language resources in the field of Natural Language Processing.

2. What a Corpus is and How to Choose One

A corpus is as substantial collection of machine readable language texts either in written form or as transcripts of spoken language. Each of the traditional first generation corpora encompasses 15 domains, as in the Brown Corpus : press, religion, popular lore, learned, fiction, science, romance, and belles-lettres, forming a total number of 500 texts, where each text consists of about 2000 words, bringing the number of the whole corpus to 1 million words (Brown Corpus).

2.1 Some examples of corpora

Paid or free corpora for registered users include the Brown Corpus, which was created in 1961 at Brown University, Rhode Island, USA, and TIMIT, standing for Texas Instruments/Massachusetts Institute of Technology, created at MIT, a database with speech samples of American English phonemically and lexically transcribed 10 sentences, each 30 seconds long spoken by 630 different male and female speakers with the aim of creating speech recognition software (LDC93s1).

Free directly available corpora include :

- The BNC (British National Corpus), which consists of about 100 million words of written and spoken English ;
- The Gutenberg Project Texts, a provider of free electronic books ;
- The LOB (Lancaster-Oslo-Bergen) Corpus, which is a replication of the Brown Corpus, also contains 1 million words of written texts ;
- The Old English Corpus, which consists of Old English texts ;
- The Bank of English, containing Cobuild Direct ; and
- The Canadian Hansard, which translates words and expressions between English and French.

A great number of different programs and search engines are available for use with corpora, which

may also be made accessible as bare texts or with annotation, ie linguistic information such as those tagged prosodically, whereby parts of the transcripts are attributed prosodic symbols, or grammatically, where parts of speech such as verb, noun, etc. are attributed (e.g., the Brown Corpus, the LOB Corpus, and the British National Corpus). These corpora can be downloaded onto a personal computer either together or as separate domains : press reportage, science fiction, or skill and hobbies, as in the Brown Corpus, for instance.

2.2 The choice of a corpus

The choice of a specific corpus depends on a variety of purposes. The first to consider is the type of language, register, or domain, one seeks to use : fiction, scientific texts, formal, informal, spoken or written language. The second important point is the use one wants to make of a corpus : a) is it for teaching grammar, vocabulary, pronunciation, or b) is it for pursuing research, for example, the historical development of some lexical terms ? The uses differ in many respects : the size of the corpus, the purpose, the time allotted to the lesson as well as the teacher's method and the availability of logistical means as well as technology-related tools among learners. For teaching grammar or vocabulary, for instance, a sample of only 30 to 40 examples is required while for a diachronic study, much larger data of language are necessary. Furthermore, written texts may be used for teaching grammar and vocabulary while the teaching of pronunciation requires appeal to a spoken corpus. A further point to take into account is the availability of corpora, as some are free and others at a cost. As regards the method, it determines the way the corpus or corpora will be presented to the learners, for example, if it is intended as a preteaching or an

inteaching activity or if the lesson will be presented deductively or inductively (for more details, see section 3.1).

3. Tools

Both online and offline tools are available for use with corpora. Online tools may be accessed via an internet connection directly without installation while offline tools must be installed on the computer. These tools allow users to create frequency wordlists, concordances, text profiles from their own texts or from web pages of their choice.

3.1 Online tools

Examples of online tools are Lextutor (Compleat Lexical Tutor) ; KWIC (Key Word In Context), a concordancer tool that allows the search for each word in a text ; and Wordsmith, a comprehensive (but with charge) package tool. These have the advantage of allowing free, direct access to a much greater number of corpora than offline programmes.

3.2 Offline tools

Offline tools are software packages that can be installed and later be used offline. One of the most popular tools is AntConc (Anthony), a software designed by Laurence Anthony, a Professor in the Faculty of Science and Engineering at Waseda University, Japan, whose areas of interest include corpus linguistics, educational technology, and natural language processing. AntConc offers a toolkit comprising frequency word lists, comparison of word lists and many functions such as the concordancers and the concordance plot . This toolkit is compatible with many computer programmes : Windows (3.5.7) ; Windows 64-bit (3.5.7) ; Macintosh OS X 10.6-10.12 (3.5.7) ; Linux 32-bit (3.5.7) ; and Linux 64-bit (3.5.7).

4. Applications of Corpus Linguistics

Because it grants teachers a way of ensuring a better quality of teaching and methods of confronting new shifts in education, corpus linguistics fulfills pedagogic and academic requirements. In research, CL may prove a valuable source of finding large amounts of data in the domain of language variation, for instance.

4.1 Corpus linguistics in education

Corpus linguistics is widely used in the educational field to guide both the teacher and the learner through their journey, as it presents several interests. There is an aboundig range of applications of corpora for teachers, who may

use them in order to teach grammar or vocabulary points or see the most frequently used syntactic constructions and expressions proper to a specific register. In the same vein, teachers' and learners' awareness is drawn to noticing styles more appropriate for use in one domain than in another. Resorting to corpora can also be useful for designing classroom activities, curriculi, and exercises. Using a deductive approach, teachers may explain a lesson by first providing the rules, and then presenting the data found in a specific corpus to confirm the rules. They may also adopt inductive instruction by presenting instances of authentic language use and then asking learners to make inferences. This inductive method promotes and enhances learning because learners can actively participate in their learning process by interacting directly with a countless number of words, expressions, sentences and longer discourse, thus gaining skills and opportunities for immediate feedback on their intuitions and constructing generalisations. More concretely, corpus linguistics provides users with a number of software tools such as AntConc, the lancsBox of the University of Lancaster course in corpus linguistics, and the online (without installation) lextutor tool, allowing them to create word lists, concordances, and word frequency lists, among other possibilities, from corpora such as those mentioned in 3.1. When one (or more) of these corpora is (are) downloaded, the software, for instance, the AntConc toolkit, may be installed whereby the basic functions of the tools are to search the corpus and display the hits of the word or combination of words typed in the box by the user. One of the tools in AntConc is KWIC (key word in context), which enables students and teachers alike to discover the different contexts in which words and expressions appear. Other options are available for the kinds of searches to make and the way the hits can be displayed. Laurence Anthony has made available several online video tutorials for users. The first one shows how to start using AntConc (Anthony, *Getting started*). After a brief overview of the background of the software, he describes each step of the process of searching a corpus in the AntConc software, for example, how to download the software ; how to load a corpus of text into the software ; and how to start analysing a corpus with the various tools inside AntConc. Once the software is installed, one can see two areas : the area of the files and on the right, the resuls area (for the different tabs for the tools). If one wishes to view a file, they click on the file view tool. For a word list, users click on the word list tool and search for a word and see how it is used in context, then they click on the concordance tool. Many online step by step tutorials are available that explain how to use the options of the software tool.

There are multifold applications of corpus linguistis in language teaching. Barlow (2002) suggests three areas in which it can be applied : syllabus design, materials development, and classroom activities. In designing a syllabus, the teacher makes decisions about the content of the course according to the learners' needs as regards the target language items to be taught (as cited in Krieger, 2003, p. 2). Corpus linguistics can also prove useful in developing materials that learners need. This may be done either by searching through a corpus or generating exercises based on authentic examples, and this offers students an opportunity to discover real features of language use.

The sample for both the language instances and the exercises must be representative and carefully selected to match the level and the learners' needs with regard to register, relevance, domain as well as mode and style. This has the merit of not only enabling foreign language students to see most accurate language instances but also helping them in distinguishing language styles appropriate in different and specific contexts and situations. For in some cases, learners' stylistic confusion is due to their exposure to language used in television programmes not always in accordance with the learning of English for academic purposes or at least major features of the target register. In this respect, teachers should draw their students' attention to discrepancies existing between different registers and domains as well as between written and spoken language. Additionally, corpora can be used in classroom activities, where, for example, students conduct analyses about specific language items and their concordances to discover by themselves and draw their own conclusions about features of real language use, where, in conformity with new educational trends, *'the teacher* [acts as] *a research facilitator rather than the more traditional imparter of knowledge' (*Krieger 2002, p. 2). As Richard Schmidt (1990) suggests, '*what language learners become conscious of - what they pay attention to, what they notice influences and in some ways determines the outcome of learning'* (as cited in Barlow 2002, p. 2).

4.2 Corpus linguistics for academic and research purposes

The requirement for integrating pedagogic and updated materials to cope with new educational approaches such as learner centredness and alternative assessment has found good company in corpus linguistics not only for its relevance in the domain of meeting learners' need for internalising various forms of language behaviour but also for its power to trigger autonomy and inductive exploration. Corpus linguistics can also be convenient for researchers in education, namely for the analysis of learner interlanguage and needs for the purpose of tracking the typical patterns found in varying degrees of language

competency, bringing tremendous contributions to designing curricula and developing teaching materials adapted to learners' levels of proficiency. Other areas of research are also particularly prone to the use of corpora for their need of large amounts of data : synchronic and diachronic studies of language, translation, sociolinguistics, language variation, cultural studies, and discourse analysis. Research in language variation may gain from corpus linguistics in its capacity for providing large amounts of data in countless fields and jargons in which language is used and in different situations, contexts and for a variety of purposes, thus tracking the development, emergence, or disappearance of some linguistic words and constructions through time.

5. Other Advantages of Corpus Linguistics

The fast growing tendency for the use of corpus linguistics in educational systems worldwide is not only a much needed data base in language teaching and

learning, but it also presents many other assets. Firstly, it may be used as a complementary tool to traditional classes, as is has been recently exemplified in blended teaching. Secondly, it is economical and ecological for its contribution in the reduction of physical presence of teachers and learners, whereby up to the third or even half the time allotted to traditional presential teaching may be reduced, in particular during the Coronavirus pandemic. This is also beneficial for the environment in lessening pollution resulting from excessive private and public transport. Thirdly, with the potential of radically changing the way teachers plan and conduct a class, corpus linguistics has a positive impact on the teacher's methodology as facilitator for their students or pupils to explore by themselves real life language through computers by being exposed to authentic language. The use of CL has the convenience of empowering non native teachers and encouraging learners independance in investigating actual usages and characteristics of various specific genres. Because it offers a tremendous asset of erasing geographical and cultural distances between teachers and learners worldwide, learning with the aid of corpora opens the doors of universal standards of modern educational methods and approaches. The added value of using corpora in a teaching/learning environment resides in its granting learners direct access to relevant data from which they can infer language rules by themselves, making a drastic change from the sometimes dull passive deductive methods of teaching, as evidence is directly and instantly obtained, though still representing a challenging task for learners, who might consider the chaotic nature of the data.

6. The Feasibility of Using Corpora in Mainstream Education

Corpus linguistics in education and academia has grown in popularity ; however, its implementation in language education has raised many reservations. Pérez-Paredes (2010) remarks that while 'most of the subjects taking part in such innovative experiences are adult university language learners with a wide array of analytical skills at their disposal',... non-university language learners [represent] the largest population of foreign language students in our modern societies' (pp. 4-5). These learners in mainstrean education are not sufficiently trained to look for relevant material ; therefore, appealing to corpora can prove a cumbersome, misleading, and time consuming task, as they are exposed to huge and chaotic amounts of data. Krieger suggests that data 'ought to be selected with the learning objectives of the class in mind, matching the purpose for learning with the corpus [as] it is the teacher's responsibility to harness a corpus by filtering the data for the students' (2002, p. 3). While Meunier discusses the 'clear divide between the exponentially growing number of publications in applied native corpus research and the introduction of corpus data in reference books and teaching materials on the one hand and everyday teaching practices on the other' (Meunier, 2010, p. 461-462), even questioning the 'representativeness' and 'relevance' of the samples (Meunier, 2010, p. 469), as 'the topics covered in most existing learner corpora are often miles away from the everyday needs of a vast majority of L2 school teachers who target the L2 for general purposes' Meunier, 2010, p. 465).

On a more practical side, the unavailability of personal computers and internet network connections among students (and, sometimes, teachers) makes it difficult, if not impossible, to follow or manage a technology-led course. In many countries, unfortunately, logistical means, local legal measures, and cultural issues present serious barriers. In other contexts, some schools cannot even cooperate in offering the logistical means for technology-based courses of this kind, many classrooms not being equipped with computer and data show rooms. In addition, some teachers are reluctant to use corpora because its use implies too much work and preparation and involves much pressure due to class time limits set in some schools.

Conclusion

Corpus linguistics, as a modern tool that teachers, learners, and researchers alike can utilise in the ambit of finding and analysing data, maximising exposure and competency in language, encompasses many benefits. With its attractiveness

for granting a large and exhaustive view of how language is used in a variety of domains, CL ensures a better quality of technology-based instruction. The positive washback of using corpora is tangible in any field of linguistics, for example, syntax, semantics, sociolinguistics, or even writing, not only for raising consciousness about authentic language use but also increasing expertise in stylistic variation and syntactic accuracy. In offering autonomy, active learning, and exposure to international standards of teaching and learning by erasing geographical and sometimes cultural and social distance, the use of corpora may also serve as a complementary ecological and economical means of teaching, learning, and conducting research. Using corpora helps learners in engaging in their own learning process for it fosters their autonomy and awakens their curiosity, thus highly upgrading their self confidence and critical thinking. This being said, and adding to the lack of experience and unfamiliarity with digital intelligences and skills required for teachers to implement corpus linguistics, the logistical means required for technoloy-based teaching and learning are often scrace and sometimes altogether lacking among large populations worldwide, in particular among children from low income families with no resources such as computers, networks, software tools, not to mention the unaffordability of the cost of corpora. This setback may be overcome by providing training in educational technology for teachers, who may resort to corpora as a preteaching task when planning lessons or designing exercises by handing out selected samples of discourse types matching the purported course objectives, thus exposing their learners to authentic language instances of various and varied registers. Until each and every student has full access to information and communications technology, it is strongly suggested that libraries and classrooms be equipped with computers and internet networks.

References

Anthony, L. (n.d.). *AntConc Homepage*. URL : http://www.laurenceanthony.net/software/antconc/

Aston, G., ed. 2001. *Learning with Corpora*. Bologna, Italy : Cooperativa Libraria Universitaria Editrice Bologna.

Compleat Lexical Tutor (n.d.). URL: https://www.lextutor.ca/conc/eng/

Higgins, J. /& Johns, T. 1984. *Computers in Language Learning*. London : Collins.

Hunston, S. 2002. *Corpora in applied linguistics*. Cambridge : Cambridge University Press.

Higgins, J. & Johns, T.,1986. "Microconcord : a language-learner's research tool". System14, 2,151 162.

Krieger, D. 2003. "Corpus linguistics : What It Is and How It Can Be Applied to Teaching". *The Internet TESL Journal for Teachers of English as a Second Language*. vol. IX, No. 3, March 2003.

http://iteslj.org/Articles/Krieger-Corpus.html
Lancaster University (n. d). Corpus Linguistics : Method, Analysis, Interpretation. https://www.futurelearn.com/courses/corpus-linguistics
Linguistic Data Consortium (LDC). https://www.ldc.upenn.edu/
McCarthy, M.J. & O'Keeffe, A. (eds). 2022. *The Routledge Handbook of Corpus Linguistics.* (2nd ed.). Routledge. https://doi.org/10.4324/9780367076399
Meunier, F. 2010. "Corpus linguistics and second/foreign language learning : exploring multiple paths". In RBLA Belo Horizonte, v.11, n.2, p. 459-477, 2011. https://www.scielo.br/j/rbla/a/BLJ6xy89SLRH7KNYSzJTPjS/?format=pdf&lang=en
Pérez-Paredes, P. 2010. "Corpus Linguistics and Language Education in Perspective : Appropriation and the Possibilities Scenario". In Harris, T., & Moreno, Jaén, Marià (eds). *Corpus Linguistics in Language Teaching.* https://www.researchgate.net/publication/234472308_Corpus_Linguistics_and_Language_Education_in_Perspective_Appropriation_and_the_Possibilities_Scenario
Sinclair, J. 1991. *Corpus, Concordance, Collocation.* Oxford : Oxford University Press.
Sinclair, J. 2003. *Reading Concordances.* London : Longman.
Sinclair, J. McH (ed.) 2004. *How to Use Corpora in Language Teaching.* Amsterdam/Philadelphia : Benjamins Publishing Company.
Schmidt, R. 1990. "Input, interaction, attention, and awareness : the case for consciousnessraising in second language teaching". Paper prepared for presentation at *Enpuli Encontro Nacional Professores Universitarios de Lengua Inglesa*, Rio de Janeiro.
TIMIT Acoustic-Phonetic Continuous Speech Corpus https://catalog.ldc.upenn.edu/LDC93s1
Tribble, C. & Jones, G. 1990. *Concordances in the Classroom.* London : Longman. University of Essex.(n.d.). *Corpus Linguistics.* https://www1.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/introduction3.htm

Abstracy

The relevance of corpus linguistics in language teaching and learning has witnessed challenging views in academia. This paper discusses the feasibility of its implementation as an educational, pedagogic, and academic requirement in language education. The claim made here is that corpus linguistics is not merely a globalisation trend, but it may prove an indispensable tool encompassing countless applications for learners, teachers, and resarchers alike even in less technology-based educational contexts.

Keywords

AntConc, autonomy, Brown corpus, corpora, corpus linguistics, ICT.

ملخص

لسانيات الكوربوس هي تخصص جديد نسبيًا يضمن دراسة وتحليل المجموعات المقروءة بالكمبيوتر (الكورپوورا)، وهي مجموعات من اللغة المكتوبة والمنطوقة في عدة مجالات تتراوح بين التقارير الإخبارية، والخطاب الأدبي، والمراجعات الصحفية، والعلوم وعدد من مجالات البحث مثل تحليل الخطاب ، وعلم اللغة الاجتماعي ، وعلم المعاجم ، وتعليم وتعلم اللغة. لقد شهدت علاقتها بتعليم اللغة وتعلمها وجهات نظر متناقضة في الأوساط الأكاديمية. يناقش هذا المقال جدوى تنفيذ استخدام الكورپورا كمتطلب تعليمي وتربوي وأكاديمي في تعليم اللغة. الادعاء هنا هو أن علم لسانيات الكوربوس ليس مجرد موضه للعولمة ، ولكنه قد يثبت أنه أداة تشمل تطبيقات لا حصرلها للمتعلمين والمعلمين والباحثين. يطرح هذاالمقال أيضًا بعض القضايا المتعلقة باستخدام الكورپوورا في سياقات تعليمية أقل اعتمادًا على التكنولوجيا

الكلمات المفتاحية

أنت كونك، التعلم الذاتي، براون كورپس، الكورپوورا، لسانيات الكورپس، (تكنلوجيا المعلوومات و الاتصالات

Résumé

La pertinence de la linguistique de corpus dans l'enseignement et l'apprentissage des langues a suscité des points de vue différents dans les milieux universitaires. Le présent article traite de la faisabilité de son implémentation en tant qu'exigence educative, pédagogique et académique dans l'enseignement des langues. L'affirmation faite dans cet article est que la linguistique de corpus n'est pas simplement une tendance mondiale mais qu'elle peut s'avérer un outil indispensable et aux innombrables applications pour les apprenants, les enseignants et les chercheurs, même dans des contextes éducatifs moins basés sur la technologie.

Mots-clés

AntConc, autonomie, corpus de Brown, linguistique de corpus, TIC.