Larisa Grčić

Department of French and Francophone Studies, University of Zadar
lgrcic@unizd.hr

Learning about Corpora, Learning with Corpora

Abstract

Integrating corpora in university education has not been an easy task and still represents a major challenge. Despite the acknowledged advantages of the use of corpora in the realms of language teaching, translation, and research, the implementation of data-driven techniques and tools at university is very slow and partial. The aim of this paper is to offer an insight into the different ways in which corpora can be exploited in a high education environment, especially for language learning and translation.

Keywords: corpora, language learning, language teaching, translation education

1. Introduction

Corpus linguistics, a field that emerged with the advent of corpora, has revolutionized linguistic research. Corpora has gained popularity in a range of language-related disciplines over the past five decades allowing a shift from an intuitive and introspective armchair approach to empirical descriptive linguistics. As a complementary resource to dictionaries and grammars, corpus data provide large and diverse authentic datasets allowing the identification but also an in-depth analysis of linguistic patterns, language variation, diachronic changes, language evolution. Corpora have become valuable resources for sociolinguistics, comparative linguistics, contrastive analysis, lexicography, language typology, forensic linguistics, media studies, gender studies, and many others. They have had a significant impact on the development of language technologies as they rely on vast amounts of linguistic data to function effectively and provide accurate results. In the age of artificial intelligence and NLP, corpora still play a crucial role in training and testing language models, chatbots, and translation algorithms (Absalom 2021).

In this paper, we attempt to present a review of applied corpus-related research that confirms the innovative use of corpora regarding language learning, language pedagogy, and translator education. Introducing corpora into university education offers an opportunity to apply an empirical method of studying language in use supporting inductive learning. Both teachers and students can benefit immensely from integrating corpora in their investigations dedicated to various aspects of

language, including morphology, syntax, phraseology, semantics, discourse analysis, pragmatics, terminology, and many others. As corpus linguistics is a wide area of research, it is impossible to present the whole range of possible uses. Therefore, our literature review has the purpose to provide basic background information for those who are not yet familiar with corpus methodology and its advantages and are interested in starting to use corpora.

The article begins with the overview of the key groundbreaking moments in the evolution of the corpus approach. Following the introduction, Section 2 is divided into three subsections providing background information about the main corpus features such as compilation criteria, different corpus types, and query techniques. In Section 3, we give a brief overview of the corpus activities aimed at developing comprehension and production skills in language learning. The contribution of corpus-based pedagogy to translator education is described in Section 4. Outlining some of these earlier experiences in corpus pedagogy, we aim to illustrate the potential of corpus use in higher education.

2. Designing and Building Corpora

Before being applied to various spheres of research, first corpora were compiled for lexicographical purposes, like the one used for *Oxford English Dictionary* published in 1928. With the arrival of computer units at universities in the 1960s, computational lexicography started developing. The largest computer corpus of the English language was created in a groundbreaking research project at the University of Birmingham led by J. Sinclair, as a support for creating a corpus-based COBUILD dictionary published in 1987. The corpus-based approach was innovative in considering language use as evidence for pre-existing linguistic theoretical statements.

A more creative vision of language was developed within a corpus-driven perspective which aimed to make theoretical hypotheses based on observation and empirical evidence. This approach was successfully incorporated within the framework of advocacy for the so-called Data-Driven Learning (DDL) introduced by Johns (1990, 1991). The basic postulate of the DDL approach was that students should acquire language knowledge inductively, using authentic data and the independent discovery of rules in them. As the learning-centred approach aligned with constructivist approaches to language acquisition, data-driven methodology was integrated not only in foreign language learning but also in LSP and CLIL classes. The direct use of corpora for teaching grammar and facilitating lexis acquisition or developing academic literacy and linguistic reflection skills was introduced at university level. However, the benefits of applying corpora were also confirmed for lower-level learners (Braun 2007; Ackerley and Coccetta 2007). Indeed, as Tognini-Bonelli pointed out, in the field of language teaching, corpus linguistics has changed "both the object to be taught and the way it is taught" (2001: 23).

Although both corpus-based and corpus-driven research have shown countless

possibilities for exploring the creativity, innovation, and potential of language phenomena, linguists emphasize that each corpus has its limitations, as no corpus can fully represent the entire language use. Some of the main issues related to corpus design are presented in the next chapter.

2.1. Corpus Compilation

While discussing methods in corpus linguistics, most frequent issues concern the corpus content, its reliability, size, and form. As these questions address fundamental concerns regarding corpus compilation, we suggest revisiting the definition of a corpus provided by Sinclair (1991): "a corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research." According to the mentioned features, the author assumes three essential principles for corpus compilation. The first refers to the selection of language samples or sampling, the second to the principle of balance, and the third to the principle of representativeness.

In corpus design, sampling implies choosing adequate texts according to the research purpose. Since no sample can be representative of the entire language system, they should be selected according to their communicative function and not according to their language features (cf. Clear 1992). In terms of size, there is no consensus on how large a text sample must be in order to be representative of a given text as a whole. However, selecting or excluding certain parts of the text for the sample can influence the research results, given that certain features can be distributed differently in the text. Size is crucial in corpus linguistics, but for certain purposes, such as LSP studies, a smaller corpus of several hundreds of thousands of words can be sufficient. The general language corpus, on the other hand, is never finite enough to be adequate for a comprehensive description. In 1963, the one-million-word *Brown Corpus* (first large computerized corpus of American English for linguistic analysis) was considered representative, while the size of the web-based contemporary *COCA Corpus* is one billion words. Besides the difference in size, the content of the two corpora is inherently different as the first is systematically assembled, while the second is a web-derived corpus. Table 1 illustrates the composition of the Brown corpus (cf. Weisser 2016 :17).

Table 1. Composition of the Brown corpus

Label	Text Category/Genre	No. of Texts
A	Press: Reportage	44
B	Press: Editorial	27
C	Press: Reviews	17
D	Religion	17

Label	Text Category/Genre	No. of Texts
E	Skills & Hobbies	36
F	Popular Lore	48
G	Belles Lettres, Biography, Essays	75
H	Miscellaneous: Government Documents, Foundation Reports, Industry Reports, College Catalogue, Industry House Organ	30
J	Learned	80
K	General Fiction	29
L	Mystery & Detective Fiction	24
M	ScienceFiction	6
N	Adventure & Western Fiction	23
P	Romance & Love Story	29
R	Humour	9

As mentioned above, collecting a corpus implies the application of certain criteria to decide as objectively as possible which texts are representative of a particular genre, time period, language variety, and the like. For this reason, it is necessary to always explicitly state the parameters for compiling the corpus, that is, the selection of representative texts, the number of texts, the size of samples, and all other parameters that determine a corpus.

In addition to sampling, the principle of balance is an important criterion for compiling the corpus as it concerns the extent to which different text types, genres, and language modalities are represented. Depending on its function, the corpus can be composed of formal (manual, poem, newspaper article, novel, textbook, script, etc.) or informal (chats, forum posts, wiki pages, tweets, blogs, comments on social networks) genres. Ideally, the representation of genres in the corpus should be equal to the representation of genres in language use. Although most general corpora are composed of written texts of various genres, the tendency of the world's national corpora is to include up to 10% of spoken language samples collected from informal conversations. In spoken and written records, the balance can also pertain to the age, gender, and origin of the author. Another dimension of balance concerns the selection of texts. The corpus compiler should decide whether to choose a significant text or author who is influential or well-known, or to make a random selection of texts and authors, or to adapt the texts to meet the linguistic criteria. A combined approach that selects from a wider range of text types has proven to be best.

Although there is no universal recommendation regarding the size, texts, genres, and language modality in which the examples for the corpus are given, scholars agree on the importance of the third principle of representativeness which implies the quality of corpus data. This principle ensures the proportionate collection of diverse sources that consider specific verbal environment (co-text), as well as the situational and cultural parameters (context) according to corpus function. The

distinction of two types of corpus representativeness suggested by Egbert et al. (2022) sheds light on this important criterion. The first type, the domain representativeness, is defined as a set of text types selected to be included in the corpus according to their variability and relevance for the domain. The second type, the linguistic distribution representativeness, suggests the appropriateness of the selected data for the specific linguistic research goal.

Since corpus evidence can be reliable only as much as the corpus, it is important to specify the corpus design criteria before compilation starts so that corpus has principled underpinning. All linguists agree on one thing—there is no single corpus that would serve all purposes, and each corpus is only an approximate sample of linguistic variety we want to explore. That is why each corpus is compiled in particular ways with a specific purpose in mind.

2.2. Different Corpora for Different Purposes

In this section, we aim to provide an overview of the main corpus types used within Corpus Studies without attempting to create an exhaustive list. As mentioned above, the corpus type and its features depend on its primary purpose.

Before the corpus era, dictionaries were considered as the main source of reference and a gold standard for language use. With the advancement of corpora, as large samples of language, witnessing empirical evidence and based on principles of balance and representativity, they were soon accepted as reference material. This is the primary function of monolingual corpora, and nowadays all major languages, as well as many minor ones, have their own reference corpora available. According to Leech (2002) a "reference corpus is designed to provide comprehensive information about the language ...] It has to be a general corpus of wide coverage of the language, and hopefully it will be treated by its user community as some kind of "standard" for the language." Reference corpora thus contrast with specialised corpora dealing with a specific field of knowledge or a web corpus that illustrates a non-standard variety as it does not follow the linguistic design criteria.

The compilation of a multilingual corpus implies selecting sources from two or more languages that are either parallel or comparable. Parallel corpora suggest an aligned resource containing original texts and their translation, and they can be unidirectional (if the sources are primarily selected in language A and their translations in language B) or bidirectional (if the original texts are also selected in language B with their translations in language A). An example of bilingual unidirectional parallel corpus is the Pavia Corpus of Film Dialogue (PCFD), consisting of the transcriptions of 12 original film dialogues and their dubbing translations (Freddi and Pavesi 2009). An example of bilingual bidirectional parallel corpus is the English-Norwegian Parallel Corpus (ENPC) designed at the University of Oslo which consists of original texts and their translations, English to Norwegian and Norwegian to English. Large-scale multilingual parallel corpora such as the JRC-Acquis,

DGT-Acquis, EUR-Lex, and Europarl corpora are released by European Union organisations (Steinberger et al. 2014). A large collection of freely available parallel corpora is available at OPUS[1] as described by Tiedemann (2012). The CLARIN[2] infrastructure provides access to 82 parallel corpora (40 bilingual and 41 multilingual corpora) suitable for comparative research as many of them are also sentence-aligned. Some of the comparable corpora offered by CLARIN are social media corpora, corpora of academic texts, parliamentary corpora ParlaMint, newspapers corpora, and historical corpora. They contain original texts in two or more languages and share similar thematic, textual, discursive, and pragmatic parameters. Comparable corpora are suitable for contrastive studies but also for educating translators in the specialized domain to acquire specific disciplinary knowledge and its terminology. For this purpose, it is often necessary to compile a targeted DIY corpus. An interesting example of comparable corpora is the International Corpus of English (ICE) which comprises 27 national and regional varieties of English. Each of the corpora is considered comparable as they all follow the common corpus design, as illustrated in Table 2.

Table 2. The design of ICE corpora[3]

SPOKEN (300)	**Dialogues (180)**	**Private (100)**	**Face-to-Face Conversations (90)**
			Phone Calls (10)
		Public (80)	Classroom Lessons (20)
			Broadcast Discussions (20)
			Broadcast Interviews (10)
			Parliamentary Debates (10)
			Legal Cross-Examinations (10)
			Business Transactions (10)

1 Accessed January 13, 2024, https://opus.nlpl.eu/corpora

2 Accessed January 17, 2024, https://www.clarin.eu

3 The precise explanation of sampling criteria aims to illustrate the comparability between corpora: "Numbers in brackets indicate the number of 2,000-word texts in each category. The texts in the corpus date from 1990 or later. The authors and speakers of the texts are aged 18 or above, were educated through the medium of English, and were either born in the country in whose corpus they are included or moved there at an early age and received their education through the medium of English in the country concerned. The corpus contains samples of speech and writing by both males and females, and it includes a wide range of age groups. The proportions, however, are not representative of the proportions in the population as a whole: women are not equally represented in professions such as politics and law, and so do not produce equal amounts of discourse in these fields. Similarly, various age groups are not equally represented among students or academic authors." Accessed February 12, 2024, https://www.ice-corpora.uzh.ch/en/design.html

SPOKEN (300)	Dialogues (180)	Private (100)	Face-to-Face Conversations (90)
	Monologues (120)	Unscripted (70)	Spontaneous Commentaries (20)
			Unscripted Speeches (30)
			Demonstrations (10)
			Legal Presentations (10)
		Scripted (50)	Broadcast News (20)
			Broadcast Talks (20)
WRITTEN (200)	Non-Printed (50)	Student Writing (20)	Student Essays (10)
			Exam Scripts (10)
		Letters (30)	Social Letters (15)
			Business Letters (15)
	Printed (150)	Academic Writing (40)	Humanities (10)
			Social Sciences (10)
			Natural Sciences (10)
			Technology (10)
		Popular Writing (40)	Humanities (10)
			Social Sciences (10)
			Natural Sciences (10)
			Technology (10)
		Reportage (20)	Press News Reports (20)
		Instructional Writing (20)	Administrative Writing (10)
			Skills / Hobbies (10)

2.3. Corpus inquiry

The way to use corpora depends on the specific research questions, but also on the possibilities that the corpus can provide. Corpus investigation software allows various quantitative and qualitative analytical searches of corpus data, such as extracting word lists according to their frequency and distribution, generating part-of-speech and semantic annotations, concordances, thesaurus, calculating n-grams and clusters, producing different visualizations of corpus data, and many others.

Concordances, as the most accessible level of corpus use, have wide application in language teaching as they allow us to examine the occurrences and behaviour of different word forms. Using the KWIC format (Key Word in Context), learners can identify a listing of node words in a specific context. J. R. Firth's (1957) catchphrase is well known: "You shall know a word by the company it keeps." John Sinclair (1991) was the first to introduce the so-called 'idiom principle', according to which every speaker has a large number of semi-prepared or preconstructed phrases at his/her disposal. Native speakers use them unconsciously, while others have to adopt them

and learn to use them. It is precisely this aspect of usage that is the most difficult when acquiring a foreign language, and by examining the concordance, students can discover the repertoire of language specific patterns. This empirical approach emphasizes the mutual connection between grammar and lexis as advocated by Stubbs (2001: 18): "It is not the words that tell you the meaning of the phrase, but the phrase tells you the meaning of the individual words in it." While corpora provide empirical evidence on word and phrase frequency across registers, concordances help explore word and phrase meanings in context. However, it is important to understand that the corpus inquiry is only the first step towards discovering different aspects of language. As Weisser (2016: 9) points out, "once we actually have extracted some relevant data from a corpus, this is rarely ever the 'final product'. Such data generally either still needs to be interpreted, filtered, or evaluated as to its usefulness, if necessary, by (re-)adjusting the search strategy or initial hypotheses and/or conclusions, or, if it's to be used for more practical purposes, such as in the creation of teaching materials or exercises, to be brought into an appropriate form."

3. Corpora in Language Teaching and Learning

In the context of the increasingly widespread use of technological advances, language corpora have become an integral part of the new subfield of language teaching, Data-Driven Learning (DDL), focused on promoting autonomy and language awareness among learners. Numerous works dedicated to the use of electronic corpora in language teaching testify to the great potential of this method. The possible ways of using language corpora in teaching both the first and second language are very diverse and can be divided into several types. In the first case, a corpus is used as a supplement to dictionaries and grammars. Numerous authors (among many others, see Bernardini 2004; Boulton 2017, 2021; Boulton and Vyatkina 2021; Meunier 2002; Vyatkina 2020; Xu 2022) single out the use of corpora as a reference material in which certain language patterns and rules can be checked. For example, corpora have transformed the way language learners acquire vocabulary by placing it on the syntagmatic level and providing learners with real-world examples of words in context where a more nuanced understanding of word usage, collocations, and idiomatic expressions is offered. Corpora also enable learners to observe part-of-speech (POS) patterns in vast amounts of authentic sentences, which aids in grasping complex grammatical rules and syntactical nuances. The new emphasis on the interrelation between grammar and lexis has led to the discovery of a wide range of recurrent language-specific patterns and networks of paradigmatic and syntagmatic relations. By providing empirical evidence on word frequency, exploring meanings in context, and offering authentic language examples for contextual learning, teachers can enhance corpus-based pedagogical grammars and underlie the phraseological approach to pedagogy (Römer 2006; Vaughan and McCarthy 2016).

Apart from exposing students to native speaker corpora, the other important use of corpora refers to the possibility of identifying common mistakes made by language learners. This kind of learner corpora presents a rich and promising field, as demonstrated in previous research (Granger and Lefer 2020; Pérez-Paredes and Mark 2022; Granger and Lefer 2023). By analysing the language data in corpora, teachers can anticipate learners' errors and provide targeted feedback, ultimately helping students correct their linguistic shortcomings. One of the biggest learner corpora is the International Corpus of Learner English (ICLE)[4] created at the University of Louvain (Granger et al. 2020). It is the result of almost 30 years of international collaboration between universities, and contains essays written by upper-intermediate to advanced learners from 16 different non-native L1 backgrounds.

For reasons of space, many other significant techniques developed for exploiting language corpora have not been mentioned, but interested readers can refer to the corpus-based pedagogy area of research that is still expanding. New areas of learners' use of corpora in Second Language Writing (SLW), such as corpus-aided writing and mobile assisted language learning (MALL), are highlighted by Schmidt (2023). Various corpus-based case studies on teaching English for Academic Purposes (EAP) writing and disciplinary writing are presented by Flowerdew (2022). Recent studies (Szudarski 2023) also bring important advances in learning collocations in second language acquisition based on corpus use.

However, despite the existence of the mentioned possibilities, as well as many theoretical works dealing with them, corpora are still rarely integrated in language learning practices. At this point, we may note that educating prospective DDL practitioners is essential. Previous findings (Breyer 2009; Ebrahimi and Faghih 2017; Leńko-Szymańska 2017; Chen et al. 2019; Lin 2019) showed the importance of implementing DDL methods and techniques in general teacher education programmes. The adoption of technology implies not only introducing user-friendly corpora tools and focusing on hands-on experience but also encouraging direct and indirect use of authentic data with emphasis on the pedagogical knowledge for exploiting corpus results. It is therefore important to incorporate practical aspects of corpora throughout the curriculum and to equip students to integrate DDL in their future teaching.

4. Corpora in Translator Education

The results of corpus linguistic studies have been applied in translation regarding particular methods (among many others, see Bowker 1998; Dash and Ramamoorthy 2019; McEnery and Wilson 1997; Olohan 2002; Sinclair 2003, 2004a) as well as in terms of the general concept of translation, with special reference to translation competence issues (Bernardini 2022; Johansson 1998; Martín 2014; Pietrzak 2015;

4 Accessed February 23, 2024, https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html

Zanettin 2014). By accessing a vast repository of authentic, representative language data, translators' education can include different browsing activities such as frequency word-listing, producing and interpreting concordances, identifying patterns, semantic preferences, linguistic nuances, and common mistranslations, extracting keywords or key terms, sorting results, leading to more accurate and contextually appropriate translations. Additionally, corpus-based approaches allow the error annotation offering valuable insights for developing reflective activities about translator strategies. Stimulating curiosity about language and searching for answers in corpora is essential as highlighted by Bernardini and Castagnoli (2008: 44): "Indeed, successful corpus work requires first and foremost an inquisitive frame of mind, a critical attitude and an ability to detect patterns, and only secondarily (some) technical skills."

The development of the field of corpus-based translations studies (CBTS) started with Mona Baker's (1993) seminal work and the creation of the first Translation English Corpus (TEC). The empirical corpus studies of translation were first mainly oriented towards the typical translation regularities, such as the study of language simplification, lexical creativity, or stylistic preferences, which led to a series of studies devoted to translation universals (Baker 1995; Laviosa 1997), the translator's style (Baker 1999), and translation norms (Kenny 1998). While looking for translation universals, independent of language pairs, interesting culture-specific aspects were discovered, which contributed to the advancement of intercultural translation studies. Very soon the field was extended by Shlesinger (1998) into corpus-based translation and interpreting studies (CTIS) and continued to grow in three directions: literary translation, translation theory, and intercultural studies. Each of them relies on monolingual or bilingual/multilingual corpora, regardless of whether they are comparable or parallel.

Translation or parallel corpora were recognized as a valuable source for understanding the cultural nuances in translation, as they are likely to reveal patterns and detect unexpected cross-linguistic equivalents. Vilceanu (2019: 1483) reports that "(...) by their very nature, translation or parallel corpora are made up of written texts belonging to a specific genre and type of discourse, which can be considered both an advantage (in the sense of allowing for in-depth analysis and interpretation of findings as guidelines for quality assurance in translation) and a disadvantage (their application is limited to certain cross-linguistic studies)." The biggest limitation of parallel corpora is that they are rarely available for the specific language pair and specific subject.

According to Tognini Bonelli (2001: 133), the translator should ideally consider evidence of both parallel and comparable corpora. The same is stated by Bernardini (2022: 493): "To fully exploit the potential of parallel and comparable corpora, these should be used together: parallel corpora (from the public domain) may provide suggestions about translator strategies and translation equivalents, while (self-made) specialized comparable corpora of non-translated target language texts may

be used to (dis)confirm the general currency of the choices made by translators." While parallel, sentence aligned corpora can be used for developing hypotheses, comparable corpora allow the possibility of testing them (through the insight into the two similar samples of L1 and L2). From the viewpoint of the translator and translation students, both types of corpora present invaluable sources for comparing and revealing not only the degree of mutual correspondence between lexical items but also general cross-linguistic similarities and differences.

5. Conclusion

In this paper, we have given a very brief insight into the advantages of corpus methodology, simply to 'set the scene', rather than to provide an extensive coverage of the multitude of corpus applications. Due to space constraints, we were unable to cover here all the variety of corpus linguistic issues, so the goal was to provide the basis for understanding how corpus investigation can be integrated into language learning and translation. More extensive coverage of the theoretical and practical issues is available in manuals like Beeby et al. (2009), Crawford and Csomay (2016), Facchinetti (2007), Hunston (2002), Ji et al. (2016); Partington (1998), Pérez-Paredes and Mark (2021), Sinclair (2003, 2004b), Stubbs (2001), Szudarski (2023), Tognini-Bonelli (2001), Weisser (2016), Zanettin (2012) among others.

As noted at the beginning of this paper, data- or corpus-driven methodology had the fundamental role in LLM training and is at the core of the main technological innovations we are witnessing. By describing some of the multiple exploitation possibilities of the corpus, we aimed to illustrate how learning to use corpora changes the way students perceive and understand languages. The evidence feedback that can be provided from corpora brings reassurance about language and makes it easy to detect cross-linguistic features as well as innacuracies and incompatibilities. The benefits of thrustworthy corpus data is even more enhanced in the age of deep learning and big data. Along with the countless opportunities of AI tools, it is vital for learners to have access to authentic and attested information for a reliable analysis to be made.

Furthermore, this enhances even more the importance of the implementation of DDL methods and techniques in general teacher training programmes to equip students with the technical and pedagogical skills needed to exploit the existing corpora and create their own corpus-based and corpus-driven material suitable for language teaching and translation tasks. By achieving a more realistic learning experience, learners will hopefully keep the curiosity towards further research into language features and develop inspiration for discovering its potential.

References

Absalom, Matthew. 2021. "Digital corpora: language teaching and learning in the age of big data." In Beaven, T. & Rosell-Aguilar, F. (Eds.) *Innovative language pedagogy report*. 97–101.

Ackerley, Katherine; Coccetta, Francesca. 2007. "Enriching language learning through a multimedia corpus." *Recall* 19(3). 351–370.

Baker, Mona. 1993. "Corpus linguistics and translation studies: Implications and applications." In Francis, G. & Tognini-Bonelli, E. (Eds.) *Text and technology: In honour of John Sinclair*. Amsterdam: Benjamins. 233–250.

Baker, Mona. 1995. "Corpora in Translation Studies: An Overview and Some Suggestions for Future Research." *Target* 7(2). 223–243. doi: 10.1075/target.7.2.03bak

Baker, Mona. 1999. "The Role of Corpora in Investigating the Linguistic Behaviour of Professional Translators." *International Journal of Corpus Linguistics* 4(2). 281–298. doi: 10.1075/ijcl.4.2.05bak

Bernardini, Silvia. 2004. "Corpora in the classroom. An overview and some reflections on future developments." In Sinclair, J. M. (Ed.) *How to use corpora in language teaching*? Amsterdam/Philadelphia: John Benjamins. 15–36.

Bernardini, Silvia; Castagnoli, Sara. 2008. "Corpora for translator education and translation practice." In Yuste, E. (Ed.) *Topics in Language Resources for Translation and Localisation*. 39–57. John Benjamins.

Bernardini, Silvia. 2022. "How to use corpora for translation?" In 0'Keeffe, A. & McCarthy, M. (Eds.) *The Routledge Handbook of Corpus Linguistics*. 485–498. doi: 10.4324/9780367076399-34

Beeby, Allison; Rodríguez Inés, Patricia; Sánchez-Gijón, Pilar. (Eds.) 2009. *Corpus Use and Translating. Corpus Use for Learning to Translate and Learning Corpus Use to Translate*. Amsterdam: Benjamins.

Boulton, Alex. 2017. "Corpora in language teaching and learning". *Language Teaching* 50(4). 483–506. doi: 10.1017/S0261444817000167

Boulton, Alex. 2021. "Research in data-driven learning." In Pérez-Paredes, P. & Mark, G. (Eds.) *Beyond the concordance: Corpora in language education*. John Benjamins. 9–34. doi: https://doi.org/10.1075/scl.102.01bou

Boulton, Alex; Vyatkina, Nina. 2021. "Thirty years of data-driven learning: Taking stock and charting new directions over time." *Language Learning & Technology*. 25(3), 66–89.

Breyer, Yvonne. 2009. "Learning and teaching with corpora: reflections by student teachers." *Computer Assisted Language Learning* 22(2). 153–172.

Bowker, Lynne. 1998. "Using Specialized Monolingual Native-Language Corpora as a Translation Resource: A Pilot Study." *Meta* 43(4). 631–651. doi: 10.7202/002134ar

Braun, Sabine. 2007. "Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora." *Recall* 19/3. 307–328.

Chen, Meilin; Flowerdew, John; Laurence, Anthony. 2019. "Introducing in-service

English language teachers to data-driven learning for academic writing." *System* 87. 102–148. doi: https://doi.org/10.1016/j.system.2019.102148

Clear, Jeremy. 1992. "Corpus sampling." In Leitner, G. (Ed.) *New directions in English language corpora*. Berlin: Mouton de Gruyter. 21–31.

Crawford, William J.; Csomay, Eniko. 2016. *Doing Corpus Linguistics*. Oxford: Routledge.

Dash, Niladri Sekhar; Ramamoorthy L. 2019. "Corpus as a Primary Resource for ELT." In Dash, N. S. & Ramamoorthy, L. (Eds.) *Utility and Application of Language Corpora*. Singapore: Springer. 91–103. doi: 10.1007/978-981-13-1801-6

Ebrahimi, Alice; Faghih, Esmail. 2017. "Integrating corpus linguistics into online language teacher education programs." *ReCALL* 29(1). 120–135.

Egbert, Jesse; Biber, Douglas; Gray, Bethany. 2022. "A Practical Framework for Corpus Representativeness." In Egbert, J., Biber, D. & Gray, B. (Eds.) *Designing and Evaluating Language Corpora*. Cambridge University Press. 52–67. doi: https://doi.org/10.1017/9781316584880.003

Facchinetti, Roberta. (Ed.). 2007. *Corpus Linguistics 25 Years on*. Amsterdam: Rodopi.

Firth, John Rupert. 1957. *Papers in Linguistics 1934-1951*. London: Oxford.

Flowerdew, Lynne. 2022. "Using corpora for writing instruction." In 0'Keeffe, A. & McCarthy, M. (Eds.) *The Routledge Handbook of corpus linguistics*. 444–457. doi: 10.4324/9780367076399-31

Freddi, Maria; Pavesi, Maria. 2009. "The Pavia Corpus of Film Dialogue: Methodology and Research Rationale." In Freddi, M. & Pavesi, M. (Eds.) *Analysing Audiovisual Dialogue. Linguistic and Translation Insights*. Bologna: CLUEB. 95–100.

Granger, Sylviane; Dupont, Maïté; Meunier, Fanny; Naets, Hubert; Paquot, Magali. 2020. *The International Corpus of Learner English. Version 3*. Louvain la-Neuve: Presses universitaires de Louvain. https://dial.uclouvain.be/pr/boreal/object/boreal:229877

Granger, Sylviane; Lefer, Marie-Aude. 2020. "The Multilingual Student Translation corpus: a resource for translation teaching and research." *Language Resources and Evaluation* 54(4). 1183–1199. doi: 10.1007/S10579-020-09485-6

Granger, Sylviane; Lefer, Marie-Aude. 2023. "Learner translation corpora." *International journal of learner corpus research* 9(1). 1–28. doi: 10.1075/ijlcr.00032.gra

Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge University Press.

Ji, Meng; Oakes, Michael; Defeng, Li; Hareide, Lidun. 2016. *Corpus Methodologies Explained. An empirical approach to translation studies*. Oxford: Routledge.

Johansson, Stig. 1998. "On the role of corpora in cross-linguistic research." In Johansson, S. & Oksefjell, S. (Eds.) *Corpora and Cross-linguistic Research*. Amsterdam and Atlanta: Rodopi.1–24.

Johns, Tim. 1990. "From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning." *CALL Austria* 10. 14–34.

Johns, Tim. 1991. "Should you be persuaded: Two samples of data-driven learning materials." In Johns, T & King, P. (Eds.) *Classroom concordancing. English*

Language Research Journal 4. 1–16.
Kenny, Dorothy. 1998. "Creatures of Habit? What Translators Usually Do with Words." *Meta* 43(4). 515–523. doi:10.7202/003302ar
Laviosa, Sara. 1997. "How Comparable Can 'Comparable Corpora' Be." *Target* 9(2). 289–319. doi:10.1075/ target.9.2.05lav
Leech, Geoffrey. 2002. *The Importance of Reference Corpora*. Invited speech at the conference of ZIO Corpus. University of the Basque Country.
Leńko-Szymańska, Agnieszka. 2017. "Training teachers in data-driven learning: Tackling the challenge." *Language Learning & Technology* 21(3). 217–241.
Lin, Ming Huei. 2019. "Becoming a DDL teacher in English grammar classes: A pilot study." *The Journal of Language Teaching and Learning* 9(1). 70–82.
Martín, Ricardo Muñoz. 2014. "Situating Translation Expertise: A Review with a Sketch of a Construct." In Schwieter, S. & Ferreira, A. (Eds.) *The Development of Translation Competence: Theories and Methodologies from Psycholinguistics and Cognitive Science*. Cambridge Scholars Publishing. 2–56.
McEnery, Tony; Wilson, Andrew. 1997. "Teaching and Language Corpora (TALC)." *ReCALL* 9(1). 5–14. doi: 10.1017/S0958344000004572
Meunier, Fanny. 2002. "The pedagogical value of native and learner corpora in EFL grammar teaching." In Granger S., Hung J. & Tyson, S. (Eds.) *Computer learner corpora, second language acquisition and foreign language teaching*. Benjamins: Amsterdam/Philadelphia. 119–142.
Olohan, Maeve. 2002. "Corpus Linguistics and Translation Studies: Interaction and Reaction." *Linguistica Antverpiensia* 1. 419–429. doi: 10.52034/LANSTTS.V1I.29
Partington, Alen. 1998. *Patterns and Meanings: Using corpora for English language research and teaching*. John Benjamins Publishing.
Pérez-Paredes, Pasqual; Mark, Géraldine. 2022. "What can corpora tell us about language learning?" In 0'Keeffe, A. & McCarthy, M. (Eds.) *The Routledge Handbook of corpus linguistics*. 313–327. doi: 10.4324/9780367076399-22
Pietrzak, Paulina 2015. "Translation competence." In Bogucki, Ł., Góźdź-Roszkowski, S. & Stalmaszczyk, P. (Eds.) *Ways to translation*. Łódź-Kraków: Łódź University Press & Jagiellonian University Press. 317–338.
Römer, Ute. 2006. "Pedagogical Applications of Corpora: Some Reflections on the Current Scope and a Wish List for Future Developments." *Zeitschrift für Anglistik und Amerikanistik* 54(2). 121–134. doi: https://doi.org/10.1515/zaa-2006-0204
Shlesinger, Miriam. 1998. "Corpus-based interpreting studies as an offshoot of corpus-based translation studies." *Meta* 43 (4). 486–493.
Schmidt, Nicole. 2023. "Unpacking second language writing teacher knowledge through corpus-based pedagogy training." *ReCALL* 35(1). 40–57. doi:10.1017/S0958344022000106
Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
Sinclair, John. 2003. *Reading Concordances*. London: Longman.

Sinclair, John. 2004a. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

Sinclair, John (Ed.). 2004b. *How to use corpora in language teaching*? Amsterdam/Philadelphia: John Benjamins.

Steinberger, Ralf; Ebrahim, Mohamed; Poulis, Alexandros; Carrasco-Benitez, Manuel; Schlüter, Patrick; Przybyszewski, Marek; Gilbro, Signe. 2014. "An overview of the European Union's highly multilingual parallel corpora." *Lang Resources & Evaluation* 48, 679–707. doi : https://doi.org/10.1007/s10579-014-9277-0

Stubbs, Michael. 2001. *Words and phrases: Corpus studies of lexical semantics.* Oxford: Blackwell.

Szudarski, Pawel. 2023. *Collocations, Corpora and Language Learning.* Cambridge University Press. doi: 10.1017/9781108992602

Tiedemann, Jorg. 2012. "Parallel data, tools and interfaces in OPUS." In *Proceedings of LREC*. Istanbul, Turkey.

Tognini-Bonelli, Elena. 2001. *Corpus linguistics at work*. John Benjamin Publishing.

Vaughan, Elaine; McCarthy, Michael. 2016. "Research in Corpora in Language Teaching and Learning." In Hinkel, E. (Ed.). *Handbook of Research of Second Language Teaching and Learning*. Oxford: Routledge. doi: 10.4324/9781315716893.CH13

Vyatkina, Nina. 2020. "Corpora as open educational resources for language teaching." *Foreign Language Annals*. 1–12. doi: 10.1111/FLAN.12464

Weisser, Martin. 2016. *Practical Corpus Linguistics: An Introduction to Corpus-Based Language Analysis*. Wiley Blackwell.

Xu, Jiajin. 2022. "A Historical Overview of Using Corpora in English Language Teaching." In Jablonkai, R. R. & Csomay, E. (Eds.) *The Routledge Handbook of Corpora and English Language Teaching and Learning.* doi: 10.4324/9781003002901-3

Zanettin, Federico. 2014. "Corpora in Translation." In House, J. (Ed.) *Translation: A Multidisciplinary Approach*. Palgrave Macmillan. 178–199. doi: 10.1057/9781137025487_10

Zanettin, Federico. 2012. *Translation-Driven Corpora. Corpus Resources for Descriptive and Applied Translation Studies*. Oxford: Routledge.