Beyond Philology No. 17/3, 2020
ISSN 1732-1220, eISSN 2451-1498

https://doi.org/10.26881/bp.2020.3.02

Corpus analysis in applied linguistics: Selected aspects

JOANNA REDZIMSKA

Received 27.04.2020,
received in revised form 16.09.2020,
accepted 1.10.2020.

Abstract

Recently, teaching and learning processes have been significantly influenced by modern technologies. Thus, the teacher's position as the only authority in the classroom has been changed into playing the role of a guide or a facilitator who should possess the knowledge and skills to use modern technologies and to freely access data. This change is particularly visible in the field of teaching and learning languages with the application of various educational platforms and software. Since this situation has been widely discussed since the 1990s, for the sake of this article only selected aspects have been taken into account. The major focus of the present article is to present language corpus analysis as a method of activating teachers and students as participants in the Data-Driven Learning (DDL) process.

Keywords

corpus analysis, DDL, activisation

Analiza korpusowa w językoznawstwie stosowanym: wybrane aspekty

Abstrakt

Rozwój technologii w znaczny sposób wpłynął na proces nauczania i uczenia się języka obcego. W konsekwencji, nauczyciel zmienił swoją pozycję z jedynego autorytetu w klasie na rolę przewodnika oraz moderatora, który powinien posiadać wiedzę i umiejętności pozwalające na wykorzystanie technologii i ogólnie dostępnych danych językowych. Widać to szczególnie w dziedzinie nauczania języków obcych, gdzie wykorzystywane są platformy i komputerowe programy edukacyjne. W związku z faktem, iż wpływ technologii na proces kształcenia opisywany jest w literaturze przedmiotu już od roku 1990, niniejszy artykuł omawia jedynie wybrane aspekty z tego zakresu. Główna uwaga poświęcona jest zagadnieniu analizy korpusowej jako metody aktywizacji nauczycieli i uczniów/ studentów poprzez proces uczenia się opartego na danych (Data-Driven Learning).

Słowa kluczowe

analiza korpusowa, DDL, aktywizacja

1. Introduction

The development of technology and the first computers paved the way for changes in all fields of research, including teaching and learning foreign languages. Thus, the traditional methods of introducing knowledge to students as well as the practice of various skills embraced the possibility of methods connected with computers, virtual reality, and free, easy language resources available for public use.

A language resource that is of core interest to this work is represented by the language corpus and teaching/learning method that is Data-Driven Learning (DDL). One of the most obvious applications of a language corpus is that it can function

as a source of knowledge about the target language's forms, use or statistics. Thus, in this respect language corpora constitute an alternative to a dictionary where the focus is mostly on meaning and possible examples where the form is used. One should also bear in mind that a language corpus as a whole always has a digital form, compared to dictionaries that traditionally have a printed form which is subsequently accompanied by a digital form. Yet, the aim of this work is to present how corpus analysis enhances language teaching and learning by offering methods and data that are not available elsewhere. However, bearing in mind the pace of the development of corpus linguistics as well as the abundance of publications connected with this field, for the sake of this article only selected aspects and corpora are further discussed. Thus, the following parts introduce a number of suggestions related to the practical application of language corpora and analysis on the basis of selected corpora for English and Polish.

2. Corpus linguistics

Although corpus linguistics has gained its position relatively recently, the origins of corpus linguistics, yet in a form different from the contemporary one, may be traced back to the 13th century (O'Keeffe and McCarthy 2010). As O'Keefe and McCarthy point out, the need for preparing wordlists and the creation of concordances were methods of Bible exegesis where scholars (mostly monks) and their students indexed the Bible hoping to find divine authorship. Another example mentioned by O'Keefe and McCarthy with reference to religious texts is the work by Anthony of Padua who first listed concordances in the Vulgate Bible. Further developments in the methods of indexing texts for wordlists and concordances were expanded on other kinds of texts, for example Shakespeare's works were annotated for concordances until the late 18th century (O'Keeffe and McCarthy 2010).

However, it is the 20th century with the advent of computers that brought about the most significant breakthrough in the corpus approach to language. The first attempts to create a machine-readable language corpus were made in the 1960s by Francis and Kučera (the Brown Corpus). Yet, with the generative approach to language at that time, their effort met with a significant amount of criticism. Generative grammar emphasizes the importance of a speaker's intuition and it concentrates on an explanatory adequacy, looking for universal language paradigms and principles. Corpus linguistics, by contrast, focuses on descriptive adequacy and examines the well-formedness and grammaticality of sentences (Meyer 2002). At the end of the 20th century, corpus linguistics gained its position and significance as a field of study and it has been acquiring greater importance ever since.

As far as the applicability of corpus linguistics is concerned, McEnery and Wilson (2011) highlight that corpus linguistics is a useful tool for identifying and characterising particular aspects of language use as well as researching these aspects from a linguistic perspective. Further the two authors (McEnery and Wilson) point out that multiple areas of linguistics derive from corpus linguistics, yet each area requires different methodology to analyse language, which has its consequence in the distinction between corpus-based and non-corpus based studies. Since corpus linguistics accounts for the complexity of language as a communicative tool with the application of interfering data (a corpus-based analysis), it stands in opposition to the generative approach whose major task is to study context-independent and most of all universal rules of language (non-corpus based studies) (Meyer 2002).

Consequently, the above-mentioned aspects raise the question of the reasons for creating different kinds of corpora. According to Renouf (2007), the three main arguments for the creation of corpora centre around the issue of science (the scientific drive for the observation and the analysis of data to test various scientific hypotheses), a pragmatic need (defined in practical

categories of the availability of data, funding and formal and technological solutions that are required for such research) and 'a fluke' (understood as an opportunity to start a new initiative that meets certain research or market demands). Moreover, Renouf (2007) mentions that the above factors highly influence both the size and the possible applications of a corpus with the tendency for small and specialised corpora, e.g. Freiburg-LOB Corpus of British English (FLOB) or the Freiburg-Brown Corpus of American English (FROWN) to compare relatively modern corpora with earlier corpora.

Thus, the application of language corpora is the most significant aspect motivated by the need for the investigation of language use in context, where the research data that is collected from a vast array of language users is the greatest benefit to the analysis (Meyer 2002). The usability of a given corpus is partially defined by its size as Meyer (2002) states that large corpora are particularly necessary for inferring details connected with grammatical constructions, forms, frequency, context or communicative power, whereas smaller corpora also possess scientific potential as long as they contain a collection of particular constructions. Undoubtedly, these are lexicographers that benefit from the use of corpus analyses by inferring information about lexical units, their range, morphological realisations and possible meanings; additionally, most of the lexicographic analysis is a largely automatic process (performed by means of computer programmes that provide data such as frequencies of words, lemmas, key words in context, tagged parts of speech) (Meyer 2002). Furthermore, the above method, as Mayer (2002) claims, is also widely applied to studying meanings and the actual uses of words which, without a corpus, are difficult to identify.

Additionally, language corpora are a way of registering language variations of different kinds, such as sociolinguistic characteristics (gender, age, ethnicity) that are represented in metadata. Following Meyer (2002), there is a choice of software

that can be used for the above purpose, an example of which is SARA (available at natcorp.ox.ac.uk/archive/SARA/index.xml).

Historical linguistics can also profit from corpus linguistics and corpus analysis. Two examples of this kind of corpora are the LOB and FLOB corpora (two parallel synchronic corpora) where one can compare language changes as well as variation in grammar and lexis (Renouf 2007). However, as Renouf (2007) points out, diachronic corpora are very often based on chronologically ordered texts or corpora that offer a selection of consequent texts (RDULES unit of the AVIATOR project available at rdues.bcu.ac.uk/aviator.shtml), which allows for the analysis of productive and creative aspects of language, collocation changes as well as word sense or meaning modifications.

Still other fields like translation studies or contrastive analysis develop due to the use of parallel corpora which (according to Meyer 2002) provide information about syntax, morphology or pragmatic aspects of translated text that can be further contrasted and compared. Parallel corpora, based on bilingual dictionaries created for this purpose, can be used for training translators and although it is a demanding task, there is software like Paraconc (paraconc.com) that facilitates the above mentioned procedures (Meyer 2002).

2.1. Examples of corpora

Corpus linguistics has gained its popularity recently, which has as its consequence the fact that a growing number of scholars and businesses are interested in projects which allow for the creation of corpora and making such corpora publicly available. As Lee (2010) points out these are not only English language corpora that are commonly used for corpus analysis but also public corpora for other languages which find their application in language study and research. The access to corpora is offered by distribution agencies and archives sites, with International Computer Archive of Modern and Medieval English (ICAME) (icame.uib.no), Linguistic Data Consortium (LDC) (ldc.upenn.

edu), CLARIN-PL (Common Language Resources and Technology Infrastructure available at clarin-pl.eu/) for Polish, and the Oxford Text Archive (OTA) (ota.bodleian.ox.ac.uk/repository/xmlui) to name a few, but as Lee (2010) highlights, access may be restricted due to the copyright or funding of these corpora.

Additionally, it must be underlined that, as far as parallel corpora are concerned, these are bidirectional and offer information about source texts as well as their translations to facilitate comparison between languages (Lee 2010). One such project that allows for the creation of lexicons and also monolingual corpora in 14 languages is The Preparatory Action for Linguistic Resources Organisation for Language Engineering (PAROLE). It offers standards and specifications for cross-linguistic analysis (Lee 2010). As far as strictly bidirectional parallel corpora are concerned, Lee mentions the English–Norwegian Parallel Corpus (ENPC) and the English–Swedish Parallel Corpus (ESPC).

An interesting example of corpora are those that include multimodal information, including speech transcripts connected with original audio or video recordings. Following Lee (2010), this allows for research into such aspects as prosody, gestures, and situated discourse to name only a few. The Scottish Corpus of Texts and Speech (SCOTS) is often quoted as an example model of this kind of corpora with its 4 million written and spoken texts (Lee 2010) as is SPOKES (http://spokes.clarin-pl.eu/) which currently contains 247,580 utterances (2,319,291 words) in transcriptions of spontaneous conversations (Pęzik 2015).

Additionally, another useful solution for gathering necessary linguistic data is offered by the almighty power of the Internet. Thus, the Web can be treated as a corpus that allows one to find relevant data. This corpus, as Lee (2010) points out, is either dynamic or static including information connected with one particular moment of use or information that is constantly updated for new language sources. Examples of this application of the Internet include Web concordancers (e.g. WebCorp, WebKWiC, KWiCFinder) to make research into concordance, the

Linguist's Search Engine which can be used to examine syntactic structures on the basis of parsed trees and the static web corpus ukWaC where two billion English words are lemmatized and tagged for parts of speech (Lee 2010).

2.2. Learner corpora

Since the major focus of the present work is on the relationship between language corpora, corpus analyses and their possible applications in language teaching and learning, it must be emphasized that these pedagogical implications resulted in the appearance of non-native speaker corpora (including written and spoken learner language). The corpus released in 2002 by Granger, Dangneaux and Meunier serves as an illustration of this pedagogical trend. In Tribble (1997) or Aston (2002) one can read about corpora created by students which centre around either genres or topics of particular interest to the group of students. Further, Braun (2005) developed a corpus of spoken English – ELISA – on the basis of a collection of interviews. Following Widdowson (1991, 2003), ELISA incorporates the principle of pedagogical mediation and the entire corpus is consistent, as far as pedagogical conceptualization is concerned, with respect to annotation, enrichment and search procedures. Thus, it promotes authentic data for learners since it uses both a great deal of decontextualized textual data as well as context-dependent interaction data (Widdowson 2003). It is worth noting that the European Minerva project SACODEYL (2005-08) (Braun 2010, Hoffstaedter and Kohn 2009, Pérez-Paredes and Alcaraz-Calero 2009, Pérez-Paredes 2010, Widmann, Kohn and Ziai 2010) also uses ELISA's pedagogical approach to a great extent including the design and corpus tools.

However, there are corpora dedicated to students who learn foreign languages. An example of such corpus is the Longman's Learner Corpus based on data gained from ESL students. Later, as Meyer (2002) points out, this corpus was used to write a dictionary which included suggestions for students' common

mistakes and strategies on how to counteract them. This information is also useful for teachers. Also, Lee (2010) references the International Corpus of Learner English (ICLE) created on the basis of students' argumentative essays illustrating different English language backgrounds.

Two further interesting examples of learner corpora are the CHILDES database and the Polytechnic of Wales (POW) Corpus (Lee 2010). These are resources that focus on data from children acquiring their native language. These resources are known as developmental corpora and they can assist in research into the way language forms are developed during the process of learning a first language (Lee 2010).

Obviously, this referential function as far as language is concerned is also fulfilled by traditional reference grammars that offer advice on how to form grammatical constructions in accordance with the rules of language (largely a prescriptive approach). An example of this is the corpus-based research of Quirk, Greenbaum, Leech, and Svartvik, which was published in 1972 (Meyer 2002). These scholars were pioneers in using corpora of written and spoken language to explain grammatical constructions.

3. Data-Driven Learning (DDL)

Data-Driven Learning (DDL) seems to be the best solution for the development of metalinguistic knowledge and learner autonomy since this method applies authentic language materials "to empower both teachers and students to develop competences in moving away from mere surface features of a text to selecting and understanding meanings and structures" (Corino and Onesti 2019: 1). One of the first advocates of this method was Johns (1991) who compared every student to Sherlock Holmes discovering the intricacies and mysteries of a language. Similarly, Sinclair (2004) praises corpus-based teaching for the use of authentic language materials. Moreover, Cobb and Boulton (2015) highlight that what is most valuable to the method is

the substantial exposure to authentic language input in a controlled way. Furthermore, among the merits of DDL, Boulton (2016: 3) emphasizes the exploitation of the following elements/aspects: authenticity, autonomy, cognitive depth, consciousness raising, constructivism, context, critical thinking, discovery learning, individualization, induction, learning-to learn, life-long learning, (meta)cognition, motivation, noticing, sensitization and transferability. However, it must be acknowledged that using DDL as an effective method requires time, practice, computer skills and most of all it must find favour with the students (especially those who do not feel comfortable with technological devices). Also, as Meunier (2011) points out, DDL necessitates considerable user investment in time and practice in order to use the data efficiently. As a result the role of a teacher changes from that of a sole authority possessing necessary knowledge to that of "a consultant, guide, coach and/or facilitator" (Suan Chong 2016). As far as students are concerned, whenever they attempt to solve language problems, they activate HOTS (higher order thinking skills), which will result in long-term knowledge retention and improved language skills (Corino and Onesti 2019: 2). Thus DDL, being a hands-on approach, provides opportunities for both teachers and students in indirect and direct applications of corpora in teaching and learning languages.

4. Discussion

As has been discussed above, there have been various types of corpora and different reasons for their creation. Without any doubt, language corpora are valuable language resources with multiple applications and the potential to fulfil different functions. However, the aim of this work is to see if corpus analysis (or working with corpora) can influence the teachers' work and facilitate or enhance the process of learning. Thus, the assumption that is made for the sake of this article is that corpus analysis is a method of activating teachers and students. As follows,

the further discussion focuses on selected aspects connected with possible practical uses of corpus analysis in the teaching/learning process.

The first and foremost aspect of corpus analysis concerns the idea of the corpus as a source of knowledge about language itself. As a result, corpus analysis allows teachers and students to have access to various kinds of language data, depending on the corpus. Some of these corpora are open-source big-data resources, for example, for English the COCA – Corpus of Contemporary American English. If a given corpus is a current project, it is updated with actual uses of language, which makes it a more reliable and applicable resource.

4.1. Teachers

Without any doubt, the most obvious, and at the same time the most significant, function of a language corpus is that it provides knowledge about a language. As has been already mentioned, the purpose of the corpus dictates what kinds of texts are used to build it and, consequentially, what kind of language forms are to be expected.

The job of teachers constantly involves various kinds of interaction with their students. Beginning with lectures and classes through to meetings with their parents, this formal, and at the same time special, relationship always relies on cooperation. There are also physical representations of this cooperation in the form of tests, essays or exercises with a twofold role: on the one hand, they are proof of the students' level of knowledge and competences and, on the other hand, at the same time they provide evidence of mistakes and issues that have to be improved. Such evidence can be collected in a form of a corpus where only language data is gathered (without any personal detail). This collection can be further used to prepare additional teaching materials to revise the problematic issues. Additionally, the frequency and quantity of certain mistakes can prove the need for further reconsideration and revision of teaching syllabuses or

even software so they will be better suited to the real needs of the students.

Another issue connected with corpus analysis is inevitably related to the question of developing a teacher's competences and activating the process of teaching and learning. Some teachers meet the challenge of building their own corpus. In a practical sense this means first learning about the programmes and tools that can be helpful in creating such corpus (developing their computer skills, learning the new software necessary to build a corpus) and then collecting texts that provide language data for the corpus (developing research skills). However, teachers who do not want to build their own corpora can use resources which are already available and look for the necessary data in them (developing analytical skills). Yet, it must be also pointed out that the most demanding task for teachers is still to give focused directions to their classes and to guide their students through data discovery and interpretation since language corpora only provide language data without any analysis. Thus, the major responsibility of teachers (and later students) is to evaluate the information found.

As follows, creating such a corpus and later analysing it seems to be a way to activate teachers, because one of the main adversaries of every teacher is routine. To avoid routine, teachers attend various courses and trainings to raise their qualifications or to look for some alternative solutions for making their lessons or courses more interesting and inspiring to their students. This results in a situation where creating and analysing their own corpora is an additional instrument which allows teachers to break up the school routine and makes their job more attractive.

4.2. Students

Corpus analysis can be profitable for students as well. Introducing corpora as an alternative to dictionaries not only broadens learners' knowledge about possible language resources but also

offers a new, technology-oriented method of learning a language. Introducing learner corpora as educational projects is a worthwhile strategy since students are more motivated to work on language that comes from their own fields of interest. The benefit here is twofold: on the one hand, the student develops his or her language skills, and on the other, the student broadens his or her knowledge about a particular domain.

Furthermore, working with corpora and carrying out a corpus analysis is focused on two major tasks. The first is centred around the creation of a corpus by students. Such a corpus can include various kinds of texts, depending on its aim. To illustrate this idea, students could build a corpus of their own mistakes and another, referential corpus that represents the correct forms. Such corpora that function as reference resources will then include either their own texts with mistakes (genuine language productions) or texts which they collect from formal/ standard resources. This is particularly useful for all kinds of revision and language drills that students can do on their own. An additional value from the perspective of a student is the fact that preparing and working with one's own corpus makes the whole process of learning highly personalized and autonomous and in consequence it allows for a significant amount of learning creativity and learning liberty.

Moreover, students can benefit from the corpus analysis by using and examining prior existing corpora to find information and solutions to their particular language problems or to find applications of selected language forms. To illustrate the above issue one can refer to a case study where a student wants to consult a corpus (which then works as an outer standard language model) to learn and understand the differences in distribution and meaning between words of nearly the same meaning. This probably is a matter of intuition for native speakers but for learners of a foreign language, it may cause problems. The examples below focus on two English words *average* and *medium* in their adjectival and nominal functions and their Polish equivalent(s) since, as far as Polish is concerned, the form in an

adjectival function is the same for both *average* and *medium*. The following examples are from COCA (www.english-corpora.org/coca), NOW Corpus (www.english-corpora.org/now) and PARARELA (http://paralela. clarin-pl.eu) and were retrieved between July and September 2020. Only two kinds of information from the corpora are being further scrutinized, namely the frequency (revealing the quantitative information) and the context (presenting qualitative information), since in the opinion of the present author these are the best and most accessible ways to show the differences between the two concepts in question.

4.3. COCA

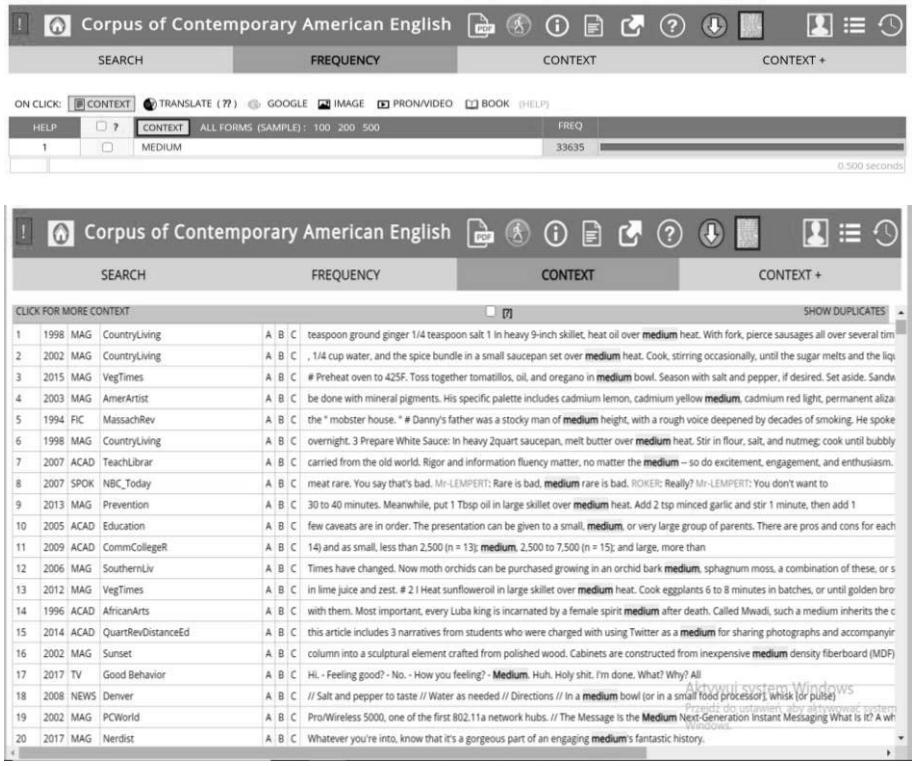

Figure 1
https://www.english-corpora.org/coca/

Upon analysing the examples above, the students find that as far as *medium* is concerned, it is used in the corpus 33,635 times. They can observe that *medium* (meaning intermediate, in-between) as a modifier is used with such concepts as size (3, 10, 18, 11), heat (1,2, 6, 9), height (5), density (16), colours (4), or mood (17) - thus such concepts whose understanding is a matter of scale or gradeability. As far as the nominal function is concerned, *medium* is used to mean 'a means, a channel of transfer' (7, 12, 14, 15, 19, 20). Tracing the examples confirms the students' intuitions and gives them an insight into the definition of the specific content of the terms.

For *average*, COCA presents the following data:

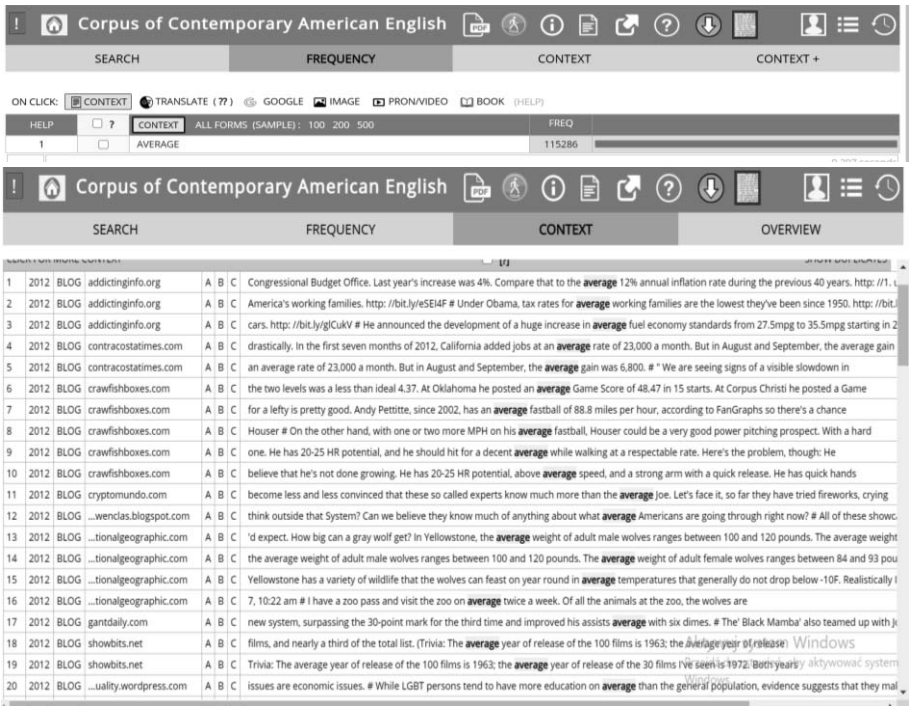

Figure 2
https://www.english-corpora.org/coca/

As has been exemplified above, *average* appears in COCA 115,286 times. Taking its function as a modifier, among 20 examples above *average* (meaning estimated on given data, approximated, being representative of) modifies such nouns as inflation (1), family (2), economy (3), fastball (7), American (12), year (19) and concepts such as rate (4), gain (5), score (6), speed (10), or weight (13). In the nominal function, *average* is used in only one example (9). Other interesting uses of *average* are represented by phrases like *the average Joe* (11) and *on average* (20).

Thus, in studying only one corpus students can see the differences between the two terms in question, in their frequency as well as in the selection of words that they are used with. So, beginning with the frequency of terms and following on to their context, the students can learn how distinct these two words are and how they should be used.

4.4. NOW

Figure 3
https://www.english-corpora.org/now/

The data above reveals that in the NOW Corpus *medium* appears 330,494 times (a number which considerably exceeds the use of *medium* in COCA). In the function of a modifier this word is used with such nouns as security (2), term (5, 22), blend (6), builds (7), business (12, 18), threat (15), or operator (16). As far as its nominal use is concerned, it is instantiated in examples 3, 4, 8, 9, 10, 11, 13, 14, 19, 20 and 21 of the above table. Thus, when compared to COCA, in the NOW Corpus (or at least in the first 20 examples) one can see that *medium* is used more as a noun than as an adjective. Additionally, the selection of nouns that *medium* modifies in NOW is different in quality from the ones that are modified in COCA, namely they are no longer nouns that require scalar modifiers.

As far as *average* is concerned, NOW offers the following data:

Figure 4
https://www.english-corpora.org/now/

According to the data in the above tables, the frequency of the word *average* in NOW is 1,704,126 times. As far as the application of *average* as an adjective is concerned, it is used with such nouns as age (26, 27), American (29), estimate (30), projection (31), escalations (33), temperature (34) or gain (35). Thus, if compared to COCA, there are two similar examples (American, gain), and the rest of examples differ. The nominal uses of the word *average* are represented in examples: 19 and 21. Yet, what draws the attention is the verbal use of *average* as provided in examples: 17, 18, 20, 22, 23, 24, 25, 28, 32 and 36, the use that has no representation in the examples from COCA. So, differences between *medium* and *average* as presented in NOW in terms of their semantic quality do not seem so obvious as in COCA. However, on the whole (and as the above analysis shows) comparing data from different corpora adds additional information for students looking to find solutions to language intricacies.

4.5. Paralela

It is highly probable that the examples described above do not provoke any questions for native speakers who, without any problems, master the qualitative differences between *medium* and *average*. Yet, these qualitative differences are the most difficult for non-native speakers who frequently look for equivalent terms in their mother tongues. Such a situation is exemplified below where *average* and *medium* have the same equivalent in Polish- *średni* (in its basic form).

Show / hide columns | Show 20 entries | First Previous 1 2 3 4 5 Next Last

Lp	English	Polish
1	If laws regulating consumer transactions were harmonised throughout the EU , small and medium businesses (SMEs) and consumers in all Member States would benefit . *Parlament Europejski*	Na harmonizacji przepisów prawa regulujących transakcje konsumenckie w całej UE skorzystałyby wszystkie małe i średnie przedsiębiorstwa (MŚP) oraz konsumenci . *European Parliament*
2	Madam President , Commissioner , just one topic should be the focus of attention when the summit is held between the European Union and Japan at the end of this month : the disaster that has devastated the people of Japan as a result of earthquakes , the tsunami and continuing radioactive contamination , and the concrete role that can be played by the EU , its Member States and individual citizens in helping to deal with the resulting problems in the short , medium and long term . *Parlament Europejski*	Pani Przewodnicząca , Panie Komisarzu ! *European Parliament*
3	Madam President , Commissioner , just one topic should be the focus of attention when the summit is held between the European Union and Japan at the end of this month : the disaster that has devastated the people of Japan as a result of earthquakes , the tsunami and continuing radioactive contamination , and the concrete role that can be played by the EU , its Member States and individual citizens in helping to deal with the resulting problems in the short , medium and long term . *Parlament Europejski*	Jeden temat powinien znaleźć się w centrum uwagi , kiedy z końcem miesiąca będzie odbywał się szczyt pomiędzy Unią Europejską a Japonią : katastrofa , która zdruzgotała Japończyków w wyniku trzęsień ziemi , tsunami i trwającego skażenia radioaktywnego oraz konkretna rola , jaką może odegrać UE , jej państwa członkowskie i poszczególni obywatele , pomagając uporać się z wynikłymi problemami w krótkim , średnim i dłuższym okresie . *European Parliament*
4	These factors will have a significant impact on energy costs in the medium term , and we will need to assess their repercussions on Europe 's current environmental strategy . *Parlament Europejski*	Te czynniki w średnim okresie znacząco wpłyną na koszty energii i będziemy musieli ocenić ich reperkusje dla obecnej strategii środowiskowej Europy . *European Parliament*

Lp	English	Polish
1	That is why , given the average level of catches over the past three years , the reference tonnage has been increased from 55 000 tonnes to 63 000 tonnes . *Parlament Europejski*	Oto dlaczego , z uwagi na średni poziom połowów w ciągu ostatnich trzech lat , tonaż referencyjny uległ podwyższeniu z 55 000 do 63 000 ton . *European Parliament*
2	I also welcome the proposal included in the report that aims to guarantee payment of the full monthly wage during maternity leave , which is 100 % of the last monthly salary or the average monthly salary . *Parlament Europejski*	Z zadowoleniem przyjmuję również zawartą w sprawozdaniu propozycję służącą zagwarantowaniu wypłacania pełnego wynagrodzenia miesięcznego w okresie urlopu macierzyńskiego , wynoszącego 100 % ostatniego wynagrodzenia miesięcznego lub średniej miesięcznych wynagrodzeń . *European Parliament*
3	Farming comes way below the average and that has got to be addressed in any future direction in which we reform the common agricultural policy , but I welcome the document . *Parlament Europejski*	Działalność rolnicza plasuje się znacznie poniżej średniej i rozwiązanie tego problemu należałoby uwzględnić niezależnie od kierunku , jaki miałaby przybrać przyszła reforma wspólnej polityki rolnej , ale z zadowoleniem przyjmuję ten dokument . *European Parliament*
4	An average Japanese girl , for example , can expect to live to the age of 83 . *Parlament Europejski*	Na przykład przewidywana długość życia dla przeciętnej japońskiej dziewczynki wynosi 83 lata , podczas gdy w Lesotho wyniosłaby 42 lata . *European Parliament*
5	It is because the World Meteorological Organisation 's figures are clear - they are average figures . *Parlament Europejski*	Ponieważ liczby podawane przez Światową Organizację Meteorologiczną są jednoznaczne - są to średnie . *European Parliament*
6	At the moment , the EU youth unemployment rate has reached 20 % , while the average EU school dropout rate is 16 % , and in some countries , like Portugal , it has reached 40 % . *Parlament Europejski*	Stopa bezrobocia wśród młodzieży sięgnęła w tej chwili w UE 20 % , natomiast średnio 16 % osób w UE przedwcześnie kończy naukę , a w niektórych krajach , jak na przykład w Portugalii , odsetek ten wyniósł 40 % . *European Parliament*

Figure 5
http://paralela.clarin-pl.eu/#search/pl/

In cases similar to the one mentioned above, an option to solve the problem of differences between apparently semantic terms is offered by corpus analysis of the original language. Furthermore, in PARALELA a student can read the different ways in which the words in question function across languages.

On the whole, the above examples of sentences from selected corpora (COCA, NOW, PARALELA) offer a wide selection of illustrations for 'language-in-use' situations for the words *average* and *medium*. However, if a student looks for real-life language applications, a reference to a corpus seems justified. Native speakers intuitively know how to use language (especially fixed expressions) in a given context. Moreover context, as a language phenomenon, has not been researched through grammar books, coursebooks or handbooks for practising 'language-in-use' situations. In other words, language learners have to learn the contextual environment for particular expressions by heart, so a reference source to check if the learners' intuition prompts a correct solution is a useful tool.

5. Conclusions

As has been discussed above, corpus analysis is a useful tool to be applied in teaching and learning foreign languages. Furthermore, selected aspects, theories and examples of corpora prove that they are valuable language resources that, on the one hand, register language forms and, on the other hand, function as reference resources available via open access to a broad public.

Yet, the main question of this article concerns the issues of how corpus analysis can influence the process of teaching and learning foreign languages. The suggestion presented above is that corpus analysis is definitely a method of activating teachers and students to enhance the educational process of teaching and learning foreign languages both inside and outside of the classroom. Furthermore, an additional advantage of using corpus analysis is the fact that students are given freedom to work on materials that they themselves identify with, as well as to

pursue their interests in selected fields which allows for a great amount of autonomy in learning.

References

Aston, Guy (2002). "The learner as corpus designer". In: Bernhard Kettemann and Georg Marko (eds.). *Teaching and Learning by Doing Corpus Analysis*. Amsterdam: Rodopi, 9–25.

Boulton, Alex (2016). "Integrating corpus tools and techniques in ESP courses". *ASP:* 69: 113-137. DOI: 10.4000/asp.4826.

Braun, Sabine (2005). "From pedagogically relevant corpora to authentic language learning contents". *ReCALL* Vol 17/1: 47–64.

Braun, Sabine (2010). "Getting past 'Groundhog Day': Spoken multimedia corpora for student-centred corpus exploitation". In: Tony Harris, María Moreno Jaén (eds.). *Corpus Linguistics in Language Teaching*. Frankfurt am Main: Peter Lang.

Cobb, Tom, Alex Boulton (2015). "Classroom applications of corpus analysis". In: Douglas Biber, Randi Reppen (eds.). *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press, 478–497.

Corino, Elisa, Christina Onesti (2019). "Data-Driven Learning: A scaffolding methodology for CLIL and LSP teaching and learning". *Frontiers in Education* 4/7. DOI: 10.3389/feduc.2019.00007.

Hoffstaedter, Petra, Kurt Kohn (2009). "Real language and relevant language learning activities: Insights from the SACODEYL project". In: Anton Kirchhofer, Jutta Schwarzkopf (eds.). *The Workings of the Anglosphere. Contributions to the Study of British and US-American Cultures*. Trier: WVT, 291–303.

Johns, Tim (1991). "From printout to handout: grammar and vocabulary teaching in the context of data-driven learning". In: Tim Johns, Philip King (eds.). *Classroom Concordancing. English Language Research Journal* 4: 27–45. DOI: 10.1017/CBO97811 39524605.014.

Lee, David, Y. W. (2010). "What corpora are available?" In: O'Keeffe, Anne, Michael McCarthy (eds.). *The Routledge Handbook of Corpus Linguistics* (1st ed.). London; New York, NY: Routledge, 107–121.

McEnery, Tony, Andrew Wilson (2011). *Corpus Linguistics: An Introduction* (2. ed., repr). Edinburgh: Edinburgh Univ. Press.

Meunier, Fanny (2011). "Corpus linguistics and second/foreign language learning: exploring multiple paths*". Revista Brasileira de Linguística Aplicada* 11: 459–477. DOI: 10.1590/S1984-63982011000200008

Meyer, Charles, F. (2002). *English Corpus Linguistics: An Introduction*. Available at https://DOI.org/10.1017/CBO9780511606311.

O'Keeffe, Anne, Michael McCarthy (2010). "What are corpora and how have they evolved?" In: Anne O'Keeffe, Michael McCarthy (eds.). *Routledge Handbook of Corpus Linguistics* (1st ed.). London; New York, NY: Routledge, 3–13.

Pérez-Paredes, Pascual, Jose, M. Alcaraz-Calero (2009). "Developing annotation solutions for online Data Driven Learning". *ReCALL*, 21/1: 55-75.

Pérez-Paredes, Pascual (2010). "Corpus linguistics and language education in perspective: Appropriation and the possibilities scenario". In: Tony Harris, Mariá Moreno Jaén (eds.). *Corpus Linguistics in Language Teaching*. Frankfurt: Peter Lang, 53–73.

Renouf, Antoinette (2007). "Corpus development 25 years on: from super-corpus to cyber-corpus". In: Roberta Facchinetti (ed.). *Corpus Linguistics 25 Years on*. Amsterdam: Rodopi, 27–49.

Sinclair, John (ed.) (2004). *How to Use Corpora in Language Teaching*. Amsterdam; Philadelphia, PA: John Benjamins.

Suan Chong, Chia (2016). "Ten innovations that have changed English language teaching". Available at https://www.britishcouncil. org/ voices-magazine/ten-innovations-have-changed-english-language-teaching.

Tribble, Chris (1997). "Improvising corpora for ELT quick-and-dirty ways of developing corpora for language teaching". In: Barbara Lewandowska-Tomaszczyk, Patrick Melia (eds.). *PALC'97 Practical Applications in Language Corpora*. Lodz: Lodz University Press, 106–17.

Widdowson, Henry, G. (1991). "The description and prescription of language". In: James Alatis (ed.). *Linguistics and Language Pedagogy: The State of the Art*. Washington, DC: Georgetown University, 11-24.

Widdowson, Henry, G. (2003). *Defining Issues in English Language Teaching*. Oxford: Oxford University Press.

Widmann, Johannes, Kurt Kohn, Ramon Ziai (2010). "The SACODEYL search tool: exploiting corpora for language learning purposes". In: Ana Frankenberg-Garcia, Lynne Flowerdew, Guy Aston (eds.). *New Trends in Teaching and Language Corpora. Proceedings of the TaLC 2008*. London: Continuum, 321-327.

Language resources

https://www.english-corpora.org/coca/
https://www.english-corpora.org/now/
http://paralela.clarin-pl.eu

Joanna Redzimska
ORCID iD: 0000-0001-7837-9397
Uniwersytet Gdański
Instytut Anglistyki i Amerykanistyki
ul. Wita Stwosza 51
80-308 Gdańsk
Poland
joanna.redzimska@ug.edu.pl