Język, Komunikacja, Informacja
I. Koutny, P. Nowak (red.)
Language, Communication, Information
I. Koutny, P. Nowak (eds.)
5/2010–2011: 37–47

Paweł Szudarski

The Role of Language Corpora in Teaching English as a Foreign Language in Poland

Abstrakt (Rola korpusu językowego w nauczaniu języka angielskiego jako języka obcego w Polsce). Badanie jest opartym na korpusie eksperymentem dotyczącym wpływu nauczania eksplicytnego na rozwój kompetencji kolokacyjnej uczniów języka angielskiego. W dwóch grupach polskich uczniów języka angielskiego jako języka obcego wprowadzono dwie różne formy nauczania: zwiększony wkład językowy (grupa pierwsza) i zwiększony wkład językowy plus ćwiczenia frazeologiczne (grupa druga). Grupa pierwsza przeczytała teksty zawierające docelowe kolokacje i wykonała ćwiczenia z zakresu słownictwa ogólnego, podczas gdy grupa druga przeczytała te same teksty, ale wykonała ćwiczenia dotyczące kolokacji. Kolokacjami docelowymi były kolokacje czasownikowo-rzeczownikowe utworzone wokół często występujących czasowników angielskich ('give', 'take', 'have', 'make', 'do') powodujące trudności w produkcji językowej w drugim języku. Trzy testy sprawdzające kompetencję kolokacyjną na różnych poziomach znajomości słownictwa ujawniły, że uczniowie w obu grupach poprawili wiedzę docelowych kolokacji i nauczanie w obu tych grupach było równie efektywne. Badanie omówione jest w kontekście wykorzystywania korpusów językowych w glottodydaktyce.

Abstract. The study is a corpus-informed experiment addressing the effects of explicit instruction on English language learners' collocational competence. Two groups of L1 Polish learners of English as a foreign language received two different forms of teaching: enriched input (the enriched group) and enriched input plus chunking practice (the enriched plus group). The enriched input group read texts containing target collocations and completed exercises focused on general vocabulary whereas the enriched plus group read the same texts but the exercises they completed were specifically focused on collocations. The target collocations were verb-noun combinations with frequent delexical English verbs ('give', 'take', 'have', 'make', 'do') known to be causing difficulty in L2 production. Three tests tapping into collocational competence at different levels of vocabulary mastery revealed that learners in both groups improved their knowledge of the target collocations and the instruction in both groups was equally effective. The study is discussed in the context of the use of corpora in teaching English and offers insights into language pedagogy.

What is corpus linguistics?

Corpus linguistics is a fast-growing methodology within linguistics in which language patterns are studied in corpora, that is large portions of authentic data (Gries 2009). According to Biber, Conrad and Reppen (1998), a language corpus is a principled collection of texts, both written and spoken, available for qualitative and quantitative analysis. Aijmer (2002) emphasizes the fact that corpora represent actual language performance showing us how language is used in different registers.

The first computer-readable corpora appeared in the 1960s. The Brown Corpus, for example, was compiled by Francis and Kucera and it contained over a million words from American English. In the 1980s, John Sinclair and his colleagues started the seminal COBUILD project based on millions of words from British English, which contributed greatly to the establishment of corpus linguistics as a new approach to language study. It needs to be remembered that collecting millions of words became possible only after technical advancements in computer science were made and large amounts of data could be stored as files. At the moment, corpora are enormous; they contain millions of words (see below) and are compiled for many natural languages. Thus, corpus linguistics as a scientific discipline is thriving offering invaluable insights into language use that have direct implications for areas such as language teaching. As McCarthy (2001: 125) maintains, corpus linguistics has altered the way we look at scientific methods and this cutting-edge change will have an influence on our notions of "education, role of teachers, the cultural context of the delivery of educational services and the mediation of theory and technique".

Spoken and written corpora

Corpus linguistics provides us with many tools for the study of language, its structure and patterns. Even a cursory analysis of the field reveals that there exist many written and spoken corpora (O'Keeffe, McCarthy and Carter 2007). One of the most widely cited corpora is the British National Corpus (BNC) that consists of 100 million words (of both written and spoken English). The written data constitute ninety per cent of the whole corpus (newspapers, books, letters, essays) and spoken data (business meetings, phone-ins, radio shows) amount only to ten per cent. There are also corpora of American English. The Corpus of Contemporary American English (COCA) is a freely available corpus that contains 410 plus million words. It was developed between 1990 and 2010 by adding 20 million words each year from both spoken and written English (spoken, fiction, magazines, newspapers, academic texts).

Compiling a spoken corpus, in comparison with a written ones, is a much more difficult task requiring recording data and carefully transcribing it. Due to this fact, as Aijmer (2002) reports, among the existing corpora written ones are more prevalent and spoken corpora are fairly small. However, even relatively small samples of spo-

ken language can cast new light on linguistic patterns. A good example of a spoken corpus is the CANCODE Corpus which stands for Cambridge and Nottingham Corpus of Discourse in English. It is a five million-word corpus of conversations recorded in everyday situations in Britain and Ireland.

Furthermore, we should also mention that there exist corpora with data collected from non-native speakers. The International Corpus of Learner English (Granger et al. 2009) contains 3.7 million words from essays written by higher intermediate to advanced learners of English from sixteen L1 backgrounds. Such databases of learner language are a rich source of information about non-native speakers whose language performance is often compared with the use of English by native speakers.

Unsurprisingly, English as a global language (Crystal 2003) dominates the field and many corpus linguists focus on English as their area of interest. However, there exist corpora with data from other languages as well. Large compilations of words from Italian, Czech, Irish and many other languages show that lexical patterning is a common characteristic of natural languages and they all equally lend themselves to linguistic study (O'Keeffe, McCarthy and Carter 2007).

How can corpora be used in English Language Teaching (ELT)?

The advent of corpora has provided new methodologies for language study and changed linguists' approach in areas such as lexicography or English Language Teaching (ELT). With regard to ELT, there are several aspects of language pedagogy which demonstrate how corpus findings have been put to practical use. First of all, all major publishers use corpus data to compile dictionaries and teaching materials. If one wishes to publish a dictionary that reflects the way English is spoken every day, they need to obtain accurate information about the language and the only way to do so is to investigate the authentic use of English by its speakers. Corpora are a reliable source of authentic language data and therefore they serve as a basis for developing a wide range of pedagogic materials such as dictionaries, coursebooks (e.g., the Touchstone series; McCarthy, McCarten and Sandiford 2005) or vocabulary books (the English Vocabulary in Use series; McCarthy and O'Dell 2002). In the past, materials developers relied on their intuition when deciding on the content of coursebooks. However, as McCarthy (1998) notes, even native speakers are inaccurate at estimating the frequency of use of different linguistic elements. Therefore, at present, when large databases of language data have become available, lexicographers and materials developers search corpora and their findings inform what is included in dictionaries and coursebooks.

It is worth mentioning that corpus analysis has led scholars to the idea of the lexical syllabus – an innovative approach to language teaching that directly uses corpus findings and organizes the content for teaching around frequent vocabulary. Sinclair and Renouf (1988) were the first authors who suggested the lexical syllabus following their work on the COBUILD project. Willis and Willis' (1988) were also interested in

the lexical syllabus and they built *the Collins COBUILD English Course* on the basis of it.

Additionally, corpora can be used by ELT practitioners as reference tools. When teachers or learners of English are unsure whether given elements are correct in terms of grammar and lexis, they can investigate their use in corpora. This is especially important for non-native teachers who are often asked by their learners why certain items in English are used 'the way they are'. Even though non-native teachers are proficient in English, it is their second language and consequently their intuitions may not be reliable. Therefore, whenever in doubt, teachers can look at corpus data and ensure that the language they teach is correct.

Furthermore, corpora often serve as large databases of language produced by learners. This enables researchers to analyze learners' linguistic development at different proficiency levels and offers useful insights into which features of English cause problems for second language learners. De Cock et al. (1998), for example, show that even at the advanced level of proficiency students misuse or underuse vocabulary, which in comparison with native speakers' performance gives the impression of 'non-nativeness' of learner English. An interesting example is the ICLE Corpus mentioned above. It is also possible to compile corpora that track linguistic development of a specific group of individuals for longer periods of time. Collecting data like this is difficult and time-consuming but such longitudinal studies help us understand how second language develops over time.

Corpus data can also be directly used in the classroom in the form of data-driven learning (DDL). This methodology was first proposed by Johns (1991: 2) who claims that the language learner "is also a research worker whose learning needs to be driven by access to linguistics data". By exploring authentic language material, learners themselves identify common patterns in grammar and lexis, while the teacher only facilitates the whole process. This inductive approach enables learners to discover which linguistic forms are used in communicative contexts and raises their awareness of how language functions in real life.

Finally, corpus linguistics can also affect how classroom-based research is conducted. It is clear that the aim of pedagogically-oriented studies is to investigate processes taking place in the classroom and consequently optimize ways in which learners are assisted on their journey to language proficiency. In light of rapid developments in corpus linguists, ELT practitioners have turned to corpora in their empirical work in order to obtain information that is only available if one has access to large databases of language data. Errors learners make are a good example here. Assuming one wants to find out what kind of grammatical errors are frequently made by advanced learners of English from a specific L1 background, it is necessary to look for such errors in a big sample of learner language in order to arrive at reliable results. While conducting research, one needs to access a representative sample of language data before any implications for teaching can be formulated. Similarly, corpora can be of great help when one selects specific language features that interest the researcher. In order to do research effectively, we need to be able to provide a rationale for why we have chosen

a given grammatical structure or a given type of vocabulary. It is also vital to be able to explain how we have chosen the items that we study. An example of such corpus-informed research is presented in the following paragraphs where a classroom study of the acquisition of collocations will be described.

Corpus-informed research – the acquisition of collocations by learners of English

The research reported below was a corpus-informed study aimed at discovering whether explicit focus on collocations can lead to improvement in learners' collocational competence at different levels of vocabulary knowledge. Specifically, the study was an attempt to examine to what extent raising learners' awareness of collocations in a post-reading activity, as advocated by Lewis in his Lexical Approach (1993), improves students' knowledge of collocations. The research design was a modification of Peters' experiment (2009) in which the acquisition of collocations by Dutch learners of English was investigated. Collocations were conceptualized as word partnerships which frequently co-occur within a given word span (Sinclair 1991).

Research questions

The pilot study reported here seeks to address the following research questions:

1. Are there gains in collocational knowledge resulting from two kinds of treatment: reading plus a general vocabulary activity (group 1) and reading plus a collocation-focused activity (group 2)?
2. Is there a difference in gains in collocation knowledge between the two treatments?

Participants

The study took place in an EFL classroom with twenty-two students of English sharing a mother tongue (Polish). Two equivalent groups of students of English were chosen: group one consisted of twelve students and group two consisted of ten students. Participants were first year university students of English philology in Wrzesnia, Poland. They had studied English for at least six years prior to the experiment (some participants for more than that) and they all passed the Matura exam, a national exam of English corresponding to the B levels of the Common European Framework of Reference for Languages (CEFR). As far as vocabulary knowledge is concerned, participants' knowledge was measured by the Vocabulary Levels Test (Schmitt, Schmitt, and Clapham 2001) at the following levels: 2000-word level, 3000-word level, 5000-word level and academic vocabulary. No significant differences between the two groups were found at these levels.

Both groups of learners followed a regular programme of study and were taught by the same teachers. Each week they had separate classes devoted to grammar, speaking, listening and reading. They also attended classes on teaching methodologies, English literature and history and all of them were conducted in English. The only difference in instruction between the two groups was the type of treatment they received from the same teacher during the experiment.

Target items

Ten verb-noun collocations of delexical verbs were chosen for this experiment. Delexical verbs, according to O'Keeffe, McCarthy and Carter (2007: 37), are a category of extremely high-frequency verbs ('do', 'make', 'take' and 'get') in their various combinations with other word classes. Other authors also include 'give', 'have', 'pay' and 'run'. For the purposes of this study, five verbs were chosen ('make', 'take', 'do', 'have' and 'give') together with nouns with which they form collocations. In all these collocations nouns carry most of the meaning of the whole phrase (e.g., 'make a proposal'; 'take a walk') and verbs become delexical.

The target collocations were selected according to several criteria. All of them consisted of individual words that were within the first 3000 most frequent words in English. Next, as far as the frequency of the whole collocations is concerned, they represented both frequent collocations (above 500 occurrences in the BNC) and less frequent collocations (below 500 occurrences in the BNC). Additionally, all target items were incongruent collocations, i.e. they coul not be easily translated from Polish into English. For example, in collocations such 'make a mistake' or 'make money', the verb 'make' is translated into Polish literally via the verb 'robić' which is a literal counterpart of 'make'. This means that such collocations are congruent in both English and Polish and therefore they were not used in the experiment. Since the items selected were incongruent collocations, the form of these collocations is realized differently in both languages. In English we 'take photos' but in Polish the same meaning is conveyed by the phrase 'robić zdjęcia' which literally means 'make photos'. Very often decoding the meaning of such collocations is not problematic for L2 learners. What causes much more difficulty is the form since it differs in learners' L1 and L2.

Treatment

The experiment took the form of the Pre-test-Treatment-Post-test design. The treatment phase lasted three weeks. It was preceded by the pre-test (administered one week before the treatment started) and followed by the post-test (administered two weeks after the treatment ended). Overall, the experiment lasted six weeks.

As far as the treatment is concerned, it was provided each week during a 90-minute lesson. Both groups received enriched exposure (input flood) to the ten target collocations that were embedded in reading texts in three consecutive weeks. Each time the reading texts were about different topics and they were followed by comprehension questions. The texts were about 1100 words long and were specifically designed for this study: the target collocations occurred twice in the text and once in the comprehension questions. The experiment lasted three weeks and overall participants were exposed to the target collocations at least nine times. It is likely that the number of exposures to the target collocations was higher since the participants attended many other classes during the experiment.

After answering the comprehension questions, both groups completed a vocabulary task. In group one, enriched group, the teacher asked participants to underline any vocabulary that they considered useful. On the other hand, participants in group two, enriched plus chunking practice group, were asked to underline any words and collocations that they considered useful. Thus, the difference in the treatment was the type of a vocabulary activity the participants were presented with. In terms of time, both groups were given the same amount to complete the vocabulary task and the same amount of teaching time was spent on underlining words and collocations. Whenever the participants had questions about the unknown vocabulary, the teacher would explain what it meant in English. No L1 translation was provided throughout the whole experiment. Finally, after the completion of the vocabulary activity, both groups discussed the topic of the reading text in pairs. This discussion lasted about ten minutes. After that, the teacher finished the lesson by summarizing the topic with the whole group.

Testing measures

Since vocabulary knowledge is a complex concept, it needs to be measured appropriately at different levels of mastery (Schmitt 2010). Therefore, in order to evaluate the effectiveness of vocabulary acquisition in both conditions, three measurement tools tapping into several aspects of collocational competence were used: a productive test of collocations in which learners were given Polish phrases and had to provide their English equivalents (Test One), a productive test of verbs in which learners had to provide verbs on the basis of definitions (Test Two) and a receptive test of collocations in which learners had to choose correct verbs from four response options (Test Three).

Results

Learners in both groups were tested twice: they took the pre-test a week before the treatment started and they took the posttest two weeks after the treatment ended. These

results of an independent-samples t-test between the pre-test results from group one and group two showed that collocational knowledge of learners in both groups before was the same (Test One: t(20)= .952 p>.05; Test Two: t(20)= .730 p>.05; Test Three: t(20)= -.736 p>.05). Therefore, any changes in collocational knowledge observed after the pre-test can only be accounted for by the effect of the treatment.

In order to answer the first research question, a series of paired-samples t-tests were conducted, comparing learners' results on the pre-test and the post-test. As shown below, there was a significant difference between learners' collocational knowledge measured on the pre-test and the post-test – both groups knew significantly more on Test Two and Test Three. On Test One, despite the fact that changes in learners' knowledge did not reach significance, the data showed a trend indicating improvement in collocational competence. These results reveal that there were gains in knowledge for both groups which means the treatment in the form of enriched input and enriched input plus chunking practice effectively contributed to the improvement of collocational knowledge.

T a b l e 1. Results of the pre-test and posttest for both groups

Test	Group one pretest (n = 12)	Group one posttest (n = 12)	PairedT-test	P value	Group two pretest (n = 10)	Group two posttest (n = 10)	Paired-samples T-test	P value
	Mean SD	Mean SD			Mean SD	Mean SD		
Test1	.14 .10	.25 .17	2.106	p = .059	.10 .10	.23 .16	2.248	p = .051
Test2	.33 .20	.52 .20	3.188	p = .009*	.28 .13	.46 .13	-4.630	p = .001*
Test3	.36 .19	.59 .16	3.102	p = .010*	.41 .12	.58 .16	3.102	p = .010*
The maximum score on all tests was 10								
The significance level was set at .05*								

The second research question concerned the effectiveness of the two kinds of treatment provided. Gains in knowledge obtained by the participants from both groups were compared but no significant differences between them were found. This means that the two kinds of treatment were equally effective in enhancing students' collocational knowledge.

T a b l e 2. Gains in collocational knowledge for both groups

Test	Group one gains (n = 12)	Group two gains (n = 10)	Paired T-test	P value
	Mean SD	Mean SD		
Test1	.11 .18	.13 .18	-.281	p > .05
Test2	.18 .20	.18 .12	.046	p > .05
Test3	.23 .17	.26 .18	.653	p > .05
The significance level was set at .05*				

Discussion

The results described above have several implications. First of all, they indicate that explicit teaching helps learners improve collocational knowledge at various levels. Secondly, the study confirms that vocabulary knowledge is a complex construct that needs to be measured at various levels (Schmitt 2010). As expected, learners seemed to have problems with collocations of delexical verbs at the productive level of their knowledge. Since we know that productive collocational knowledge is the most difficult one to acquire, it is necessary to conduct more research in order to determine which forms of explicit instruction are the most effective. Additionally, both groups of learners seemed to have improved their collocational competence in the same way. A useful follow-up would be to carry out a similar experiment on other types of collocations and with learners at different proficiency levels and from different L1 backgrounds. Additionally, more research is warranted on the issue of how to best present collocations in pedagogic materials.

Implications of corpus linguistics for language pedagogy

The study described above is an example of a corpus-informed experiment conducted in the classroom context. This investigation was focused on lexis but empirical work on many other linguistic features can also be informed by corpus evidence. Hopefully, the research reported here shows the usefulness of insights from corpus linguistics. Moreover, it is worth stressing that there exist websites which offer corpus-based applications and help conduct research. For instance, the Compleat Lexical Tutor website (http://www.lextutor.ca/) developed by Tom Cobb offers a wide range of tools that facilitate empirical investigations into language. The website enables to create frequency lists from texts or prepare cloze tests that can be administered in the classroom. Both teachers and researchers can access such web-based tools since they help conduct regular language classes as well as design sophisticated research studies. Moreover, it is noteworthy that there are websites that offer corpus-based language tasks that can immediately be used for pedagogic purposes (e.g., www.cambridge.org/gb/elt/).

As already mentioned, all major publishers collect their own corpora that are used in the development of dictionaries and pedagogic materials. Coxhead's (2000) Academic Word List (AWL) is a perfect example of this. In her corpus, Coxhead analyzed 3.5 million words from different academic disciplines, which enabled her to arrive at a list of 570 word families most needed to study at university. This research has been so influential that several materials developers have published books that aim at the mastery of the AWL (e.g., Schmitt and Schmitt 2005). Many corpus-informed ELT materials are published at the moment (Reppen 2010) and it is expected that their value as pedagogic tools will increase in the future.

Finally, as technological advancements are rapid, more corpora of different kinds are likely to be compiled. Some scholars have advocated the use of the Internet as a corpus (e.g., Keller and Rapata 2003). Others started developing multimodal corpora in which text is accompanied by video recordings (e.g., the French Corpus of Interactional Data). Thanks to that, all the information that is conveyed through modalities such as gesture or body posture can be captured as well. In addition, corpora of languages for specific purposes (e.g., corpora of business English) have become available. Undoubtedly, the way businessmen use language differs from interactions in other contexts and this information can be directly used in the development of pedagogic materials. Furthermore, corpora comprised of coursbooks and other language practice books are slowly being compiled (Meunier and Gouverneur 2009). This should help us find better ways of presenting language input to learners. Therefore, it seems fair to say corpus linguistics is well-established as a methodology for language study and more large databases with language data are likely to appear in the future.

REFERENCES

Aijmer, Karin 2002: *English discourse particles. Evidence from a corpus*. Amsterdam: John Benjamins.

Biber, Douglas, Conrad, Susan and Reppen, Randi 1998: *Corpus linguistics: investigating language structure and use.* Cambridge: CUP.

Coxhead, Averil 2000: A New Academic Word List. *TESOL Quarterly* 34(2), 213–238.

Crystal, David 2003: *English as a global language.* Cambridge: CUP.

DeCock, Sylvie, Granger, Sylviane, Leech, Geoffrey and McEnery, Tony 1998: *An automated approach to the phrasicon of EFL learners.* In Granger, Sylviane (ed.). *Learner English on computer.* London: Addison Wesley Longman.

Granger, Sylviane, Dagneaux, Estelle, Meunier, Fanny and Paquot, Magali 2009: *International Corpus of Learner English v2.* Louvain-la-Neuve: Presses universitaires de Louvain.

Gries, Stefan 2009, What is corpus linguistics? *Language and Linguistics Compass* 3, 1–17.

Johns, Tim 1991, Should you be persuaded: two samples of data-driven learning materials. *English Language Research Journal* 4, 1–16.

Keller, Frank and Lapata, Mirella 2003: Using the Web to Obtain Frequencies for Unseen Bigrams. *Computational Linguistics* 29(3), 459–484.

Lewis, Michael 1993: *The lexical approach. The state of ELT and a way forward.* Hove UK: Language Teaching Publications.

McCarthy, Michael 1998: *Spoken language and applied linguistics*. Cambridge: CUP.

McCarthy, Michael 2001: *Issues In Applied Linguistics*. Cambridge: CUP.

McCarthy, Michael and O'Dell, Felicity 2004: *English vocabulary in use.* Cambridge: CUP.

McCarthy, Michael, McCarten, Jeanne and Sandiford, Helen 2005: *Touchstone.* Cambridge: CUP.

Meunier, Fanny and Gouverneur, Celine 2009: *New types of corpora for new educational challenges: collecting, annotating and exploiting a corpus of textbook material*. In Aijmer, Karin (ed.). *Corpora and language teaching*. Amsterdam: John Benjamins.

O'Keeffe, Anne, McCarthy, Michael & Carter, Ronald 2007: *From corpus to classroom. Language use and language teaching.* Cambridge: CUP.

Peters, Elke 2009: *Learning collocations through attention-drawing techniques: a qualitative and quantitative analysis.* In Barfield, Andy and Gyllstad, Henrik (eds.). *Researching collocations in another language.* Houndmills: Palgrave Macmillan.

Reppen, Randi 2010: *Using corpora in the language classroom.* Cambridge: CUP.
Sinclair, John 1991: *Corpus, concordance, collocation.* Oxford: OUP.
Sinclair, John & Renouf, Antoinette 1988: *A lexical syllabus for language learning.* In Carter, Ronald & McCarthy, Michael (eds.). *Vocabulary and language teaching*. London: Longman.
Schmitt, Norbert 2010: *Researching vocabulary: a vocabulary research manual.* Basingstoke: Palgrave.
Schmitt, Diane and Schmitt, Norbert 2004: *Focus on Vocabulary: Mastering the Academic Word List*. White Plains, NY: Longman.
Schmitt, Norbert, Schmitt, Diane and Clapham, Caroline 2001: Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing* 18(1), 55–88.
Willis, Dave and Willis, Jane 1988: *Collins COBUILD English Course.* London: Collins.