RiCL Research in Corpus Linguistics

Introduction: the use of corpora for language teaching and learning

Antonio Vicente Casas-Pedrosa[*], Jesús Fernández-Domínguez[**] and Alejandro Alcaraz-Sintes*
Universidad de Jaén[*], Universitat de València[**] / Spain

Abstract – This paper aims at contextualizing and presenting the first volume of the journal *Research in Corpus Linguistics* and is, therefore, divided into two main parts. First of all, it provides an introduction to the field of corpus linguistics and its increasingly relevant role in language teaching and learning. Secondly, it briefly introduces and discusses the six articles of the volume. Stemming from oral presentations delivered at the 4th International Conference on Corpus Linguistics (CILC2012, Jaén, Spain), these articles have a number of features in common. They all make extensive use of corpora and at the same time deal with language teaching and language learning, the underlying assumption being that a genuine and mutually beneficial connection can be established between teaching and research. For this reason, each of them constitutes an illustrative sample of how different corpora can be exploited for different research purposes.

Keywords – corpus linguistics, language description, language learning, language teaching

This first volume of the journal *Research in Corpus Linguistics* is a special issue with six contributions that were originally presented at *CILC2012* (*4th International Conference on Corpus Linguistics*), held at the University of Jaén (Spain) from 22nd to 24th March 2012. Although these papers were presented in five of the nine thematic panels established by *AELINCO* (*Spanish Association for Corpus Linguistics*) for its conferences, they all share a common feature: they explore the use of corpora for the teaching and learning of a language.

Formerly, the main approach to the description or teaching of a language used to be prescriptive, with an emphasis on the language officially accepted as 'grammatical', while 'ungrammatical' (or non-acceptable) productions were condemned. Later developments underlined that many of those linguistic prescriptions were subjective and not based on empirical evidence, which meant a readjustment whereby the study of linguistic phenomena should be tackled from a strictly scientific point of view (Quirk et al. 1985: 14; Nelson 2005). This subsequently involved a different approach to the analysis of language, and more recent trends of language description have adopted this viewpoint. According to the new perspective, the linguist's task is the analysis of genuine language from a descriptive and empirical point of view. It is at this point that the term 'corpus' comes into play.

According to the *Oxford English Dictionary* (s.v. *corpus* 3.b), the first recorded written example of the word *corpus*, understood as "[t]he body of written or spoken material upon which a linguistic analysis is based", dates back to 1956. It appears in the following excerpt by Allen (1956: 128): "The analysis here presented is based on the speech of a single informant (…) and in particular upon a corpus of material, of which a large proportion was narrative, derived from approximately 100 hours of listening". This involves not only corpora as tools of their own, but also their use and application to the creation of other resources.

Until very recently, dictionaries, grammars and textbooks, among others, have incorporated custom-made examples drawn up by lexicographers and scholars for the purpose of illustrating, *a posteriori*, the concepts, definitions and explanations being presented. Of course, these examples were not necessarily unreliable, since they were most often made up by native speakers, who were considered "a living dictionary" (Rundell 2007: vii). These samples of language, however, lacked something that is considered essential today: the support of empirical evidence. They were often

perceived as somehow artificial, since they seemed to have been created on the spot and without a specific communicative context. Nowadays, most works on language description (teaching manuals, grammar books, dictionaries, etc.) underline the fact that they have used a language corpus as the source of examples. For example, Carter and McCarthy's (2006) backcover includes the catch-phrases "real English guarantee" and "real everyday usage", thus highlighting the idea that the description within the work is based on the actual usage – written and spoken – of the language by native speakers from all over the world.

Thanks to the advances of computer science, it is now possible to access resources essential for the linguistic description of language, first and foremost language corpora, which are also playing an increasingly relevant role in language learning and teaching. Despite their relatively short existence, corpora have already become a crucial tool for the analysis of languages, and a dynamic relationship has flourished between corpora and language teaching. The increasing number of articles, books and journals on the topic published every year, on the one hand, and the amount of conferences, symposia and workshops organized worldwide, as well as the creation of academic associations closely related to corpora and their exploitation, all bear witness to the phenomenon (e.g., Granger et al. 2002; Aston et al. 2004; Gavioli 2005; Braun et al. 2006; Hidalgo et al. 2007, among many others).

The applications of corpora for language teaching have been discussed, for example, by Leech (1997), Römer (2008) and McEnery and Xiao (2011), who differentiate between indirect and direct uses. Corpora are being indirectly used, for example, for the design of teaching syllabi with an emphasis on communicative competence, e.g., the *Collins COBUILD English Course* (Willis and Willis 1989; see Hernández this issue) or when representing the frequency of occurrence of language items in grammar and usage handbooks. Other indirect applications are found in Language for Specific Purposes (LSP) corpora, learner corpora and translation corpora, each with different implications for the language classroom. The LSP lesson, as a token, can benefit from the creation of genre-specific corpus-derived glossaries or from concordance data when creating authentic teaching materials. On the other hand, learner and translation corpora are two of the most widely employed upshots of corpus linguistics for language pedagogy, and offer a variety of practical uses, such as learner dictionaries, syllabus design or the creation of teaching materials based on error analysis (see Granger 2002; Meunier 2002; Carrió Pastor and Mestre Mestre this issue; Wierszycka this issue). As regards the direct applications of corpora, scholars have often reported positive results when students are faced with hands-on tools such as online corpora or when they are able to retrieve and discuss concordance lines on a relevant topic (Bernardini 2002; Milojkovic this issue). It seems that this kind of data-driven learning furthers an autonomous and interactive kind of learning between students and language data, while teachers are able to move from the role of information provider to that of facilitator.

Notwithstanding the above benefits, the use of corpora in the classroom is not without difficulties. Once the decision is made to employ a corpus, one of the most frequent dilemmas is whether to exploit one of the many corpora available or to compile a new one, which naturally depends on issues such as the target audience and the availability of appropriate texts. There are numerous general-purpose corpora today, especially for languages like English, French, German, Italian or Spanish, and there exists a vast range of alternatives to choose from. For more specific registers, the option of an *ad hoc* corpus is at hand as well, although such relevant issues as balance, representativeness, design or sampling will have to be seriously considered in the compilation process. It should also be remembered that large size is not always advantageous, and that smaller corpora can prove very effective if accurately selected (Leech 1997: 22; Meunier 2002: 129).

The present collection of articles intends to illustrate some of the aforementioned matters by resorting to the experiences of six scholars in this field. Following this introduction, Hernández delves into the relevance of corpus linguistics for the linguistics curriculum by looking at material from a corpus made up of computer-mediated texts. Based on the experience gathered from a research project at the University of Duisburg-Essen, this article pays attention to the language of 'digital discourse' (Crystal 2010) as found in e-mails, chat rooms, text messages, blogs, forums as well as in a variety of social media services (*Facebook*, *Twitter*, *YouTube*, etc). The article discusses the various specific procedures that these types of texts require before they can be effectively incorporated to a corpus, among them user privacy, a definition of textual units, the tagging of non-standard spellings and errors, and the codification of images or emoticons. The author also explores the possibility of implementing major issues in corpus construction into the academic curriculum in the form of project-based learning. Hernández ends by presenting a variety of new challenges and possible solutions regarding the compilation and processing of this corpus, for example, the fact that the students themselves wrote a corpus manual with a general description of the textual mark-up and processing guidelines.

In their contribution, Carrió Pastor and Mestre Mestre view errors as a key feature of language learning and focus on the identification and classification of errors related to the students' grammar acquisition process and pragmatic competence. In particular, they look at errors detected in writing with the aim of shedding light on the nature of the mechanisms that foreign learners employ in language production. In contrast to the traditional focus on grammatical errors in second language teaching, this work turns to pragmatics, here understood as the difference between the *official* meaning of a word or sentence and the meaning perceived by the hearer derived from what the speaker said (see Archer et al. 2012). From the comparison between grammar and pragmatics, two objectives are pursued here: first, a proposal for tagging grammatical and pragmatic errors according to the competences laid out in the *Common European Framework of Reference for Languages* (*CEFR*; see Council of Europe 2001) and, second, the establishment of a

correspondence between these two types of errors. The basis for this experiment is a corpus of written texts produced by undergraduate students (B1 level) at the Universitat Politècnica de València, in which the assignments were specifically targeted at the development of pragmatic and grammatical competences. The conclusions of the study are that some grammatical and pragmatic errors coincide and that such a correspondence should be taken into account by language teachers in order to help students in their language learning process.

Wierszycka concentrates on learner English as well, in this case by delving into the semantics of phrasal verbs (PVs). In view of the alleged difficulties that non-native speakers of English experience when faced with PVs (Celce-Murcia and Larsen-Freeman 1999), this article starts from the hypothesis that Polish learners of English master a significantly smaller range of PVs than English native speakers, and that their degree of use of the semantic categories of PVs is inversely proportional to the PVs' level of idiomaticity (see Dagut and Laufer 1985). The evidence for this study is drawn from the Polish component of *LINDSEI* (*Louvain International Database of Spoken English Interlanguage*; see Gilquin et al. 2010) and from the *LOCNEC* (*Louvain Corpus of Native English Conversation*; see De Cock 2003), and is framed within Granger's (1996) scheme of Contrastive Interlanguage Analysis. This comparison of PV usage confirms that while native speakers use PVs in a linear manner, considerable underuse is found on the part of Polish learners; in particular, and thanks to a previous analysis of PV compositionality, Wierszycka verifies that idiomatically opaque PVs are especially neglected when used by Polish learners.

Next, Lee examines the use of modal verbs in academic written feedback as hedging expressions (see Salager-Meyer 2011: 35). By using a corpus of around 36,000 words collected at two Humanities departments in the UK, this investigation turns its attention to the language used by tutors when giving feedback to their students. The author sets out from a wordlist of nine modal verbs (*can*, *could*, *may*, *might*, *must*, *shall*, *should*, *will* and *would*) and provides extensive evidence (frequencies of occurrence) to discuss the functions of modal verbs in the genre of written feedback, among which we find criticism, suggestion, possibility, necessity, certainties, permission and advice. Among other results, this piece of research shows that *could*, *might* and *would* are the most widely used modal verbs of the set, while *shall* is not present given its lower relevance in the context of written feedback. Intermediate positions are occupied by *may*, *must*, *should* and *will*, each with a different level of certainty that directly affects its higher or lower usage. Interestingly, the author shows that the language used by tutors is rather assertive and direct when the feedback concerns an aspect of writing, while it becomes more indirect and tentative when levelling criticism. These findings can be exploited, for instance, for the improvement of feedback-writing practices in teacher training programmes.

Milojkovic's article is a corpus-based approach to Bill Louw's (1993 and subsequent publications) Contextual Prosodic Theory, which reports the experience resulting from the application of stylistics research on the part of second-year students of English. Two research questions are posed here: one, whether text corpora can help infer authorial text, as postulated in Louw's "text reads text" principle, and two, whether this methodology can be effectively applied to the stylistics classroom. Once the foundations are laid through quantitative and qualitative tests, the experiment depicts the process of meaning construal, where the participants, after an absolute minimum of theoretical background, are provided with concordance lines as a means of interpreting a collocation in a given short excerpt. This virtual absence of instruction in principle leads to unbiased acceptance on the part of the majority of the students. The subjects are tested on semantic prosodies, absent collocates and auras of grammatical strings, through tasks that vary in their format. The relevance of this study lies in the fact that, besides supplying positive answer for the aforementioned two research questions, it is the first application of Louw's theory in the classroom, sporadic studies on semantic prosody aside. The article confirms that text does read text for the non-native students of English at the Belgrade English Department regardless of their level of proficiency.

The volume closes with Szabó's inspection of language ideologies in Hungarian school metalanguage. Revolving around an array of theoretical frameworks (Coulter 2005; Laihonen 2008; Aro 2012), this contribution draws on the *Corpus of Hungarian School Metalanguage-Interview Corpus* (*CHSM-IC*), an annotated transcription of spoken metalanguage based on semi-structured research interviews of Hungarian students, in order to investigate interactional routines used in metadiscourses. On the basis of this material, the author compares texts from well-known handbooks with interview data from *CHSM-IC* and then contrasts the participants' narratives with their own communicational experiences. The article also includes a case study on the Hungarian discourse marker *hát* ('so', 'well'), which illustrates the conflict between language ideologies disseminated by the Hungarian school system and the linguistic self-representation in the interviewees' narratives. This analysis reveals that the narratives of both teachers and students carry a negative evaluation of *hát*, and a detailed discussion of the topic concludes that the use of this marker is an important part of everyday communication practice. Accordingly, one conclusion is that metalinguistic utterances (e.g., answers on grammaticality, statements on linguistic accuracy, etc.) and observable, spontaneous (or semi-spontaneous) language use patterns are regularly not in accordance with each other.

REFERENCES

Allen, William S. 1956. Structure and system in the Abaza verbal complex. *Transactions of the Philological Society* 55: 127–176.

Archer, Dawn, Karin Aijmer and Anne Wichmann. 2012. *Pragmatics. An advanced resource book for students*. London: Routledge.

Aro, Mari. 2012. Effects of authority: voicescapes in children's beliefs about the learning of English. *International Journal of Applied Linguistics* 22/3: 331–346.

Aston, Guy, Silvia Bernardini and Dominic Stewart (eds.). 2004. *Corpora and language learners*. Amsterdam: John Benjamins.

Bernardini, Silvia. 2002. Exploring new directions for discovery learning. In Bernhard Kettemann and Georg Marko (eds.), *Teaching and learning by doing corpus analysis. Proceedings of the Fourth International Conference on Teaching and Language Corpora, Graz 19–24 July, 2000*. Amsterdam: Rodopi, 165–182.

Braun, Sabine, Kurt Kohn and Joybrato Mukherjee (eds.). 2006. *Corpus technology and language pedagogy*. Frankfurt: Peter Lang.

Carter, Ronald and Michael McCarthy. 2006. *Cambridge grammar of English. A comprehensive guide. Spoken and written English grammar and usage*. Cambridge: Cambridge University Press.

Celce-Murcia, Marianne and Diane Larsen-Freeman. 1999. *The grammar book: an ESL/EFL teacher's course*. Second edition. Boston, MA: Heinle and Heinle.

CHSM-IC = Magyar Iskolai Metanyelvi Korusz – Interjúkorpusz / Corpus of Hungarian School Metalanguage – Interview Corpus. Budapest: Research Institute for Linguistics of the Hungarian Academy of Sciences. Project chair: T. P. Szabó. <http://metashare.nytud.hu/repository/search> (17 August 2013).

Coulter, Jeff. 2005. Language without mind. In Hedwig te Molder and Jonathan Potter (eds.), *Conversation and cognition*. Cambridge: Cambridge University Press, 79–92.

Council of Europe. 2001. *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.

Crystal, David. 2010. The changing nature of text: a linguistic perspective. In Wido van Peursen, Ernst D. Thoutenhoofd and Adriaan van der Weel (eds.), *Text comparison and digital creativity*. Leiden: Brill, 229–251.

Dagut, Menachem and Batia Laufer. 1985. Avoidance of phrasal verbs – a case for contrastive analysis. *Studies in Second Language Acquisition* 7/1: 73–79.

De Cock, Sylvie. 2003. *Recurrent sequences of words in native speaker and advanced learner spoken and written English*. PhD thesis. Louvain: Université Catholique de Louvain.

Gavioli, Laura. 2005. *Exploring corpora for ESP learning*. Amsterdam: John Benjamins.

Gilquin, Gaëtanelle, Sylvie De Cock and Sylviane Granger (comp.). 2010. *Louvain International Database of Spoken English Interlanguage* (*LINDSEI*). Louvain-la-Neuve: UCL Presses universitaires de Louvain.

Granger, Sylviane. 1996. From CA to CIA and back: an integrated approach to computerized bilingual and learner corpora. In Karin Aijmer, Bengt Altenberg and Mats Johansson (eds.), *Languages in contrast. Textbased cross-linguistic studies*. Lund: Lund University Press, 37–51.

Granger, Sylviane. 2002. A bird's-eye view of learner corpus research. In Granger et al. (eds.), 3–33.

Granger, Sylviane, Joseph Hung and Stephanie Petch-Tyson (eds.). 2002. *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins.

Hidalgo, Encarnación, Luis Quereda and Juan Santana (eds.). 2007. *Corpora in the foreign language classroom*. Amsterdam: Rodopi.

Laihonen, Petteri. 2008. Language ideologies in interviews: a conversation analysis approach. *Journal of Sociolinguistics* 12/5: 668–693.

Leech, Geoffrey. 1997. Teaching and language corpora: a convergence. In Anne Wichmann, Steven Fligelstone, Tony McEnery and Gerry Knowles (eds.), *Teaching and language corpora*. London: Longman, 1–23.

Louw, William E. 1993. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In Mona Baker, Gill Francis and Elena Tognini-Bonelli (eds.), *Text and technology. In honour of John Sinclair*. Amsterdam: John Benjamins, 157–176.

McEnery, Tony and Richard Xiao. 2011. What corpora can offer in language teaching and learning. In Eli Hinkel (ed.), *Handbook of research in second language teaching and learning*. London: Routledge, 364–380.

Meunier, Fanny. 2002. The pedagogical value of native and learner corpora in EFL grammar teaching. In: Granger et al. (eds.), 119–141.

Nelson, Gerald. 2005. Description and prescription. In Keith Brown (ed.), *Encyclopedia of language and linguistics*. Oxford: Elsevier, 460–465.

Oxford English Dictionary online. 2013. Oxford: Oxford University Press. <http://www.oed.com> (30 July 2013).

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

Römer, Ute. 2008. Corpora and language teaching. In Anke Lüdeling and Merja Kytö (eds.), *Corpus linguistics. An international handbook*. Volume 1. Berlin: Mouton de Gruyter, 112–130.
Rundell, Michael (ed.). 2007. *Macmillan English dictionary for advanced learners*. Second edition. Oxford: Macmillan.
Salager-Meyer, Françoise. 2011. Scientific discourse and contrastive linguistics: hedging. *European Science Editing* 37/2: 35–37.
Willis, Dave and Jane Willis. 1989. *Collins COBUILD English course*. London: HarperCollins.