ACTAS

I Congreso Internacional de Enseñanza de Inglés en Centros Educativos

CEU | Ediciones

unimel EDUCACIÓN

INTERNATIONAL CONFERENCE
EICE 2016
ENGLISH TEACHING IN EDUCATIONAL INSTITUTIONS

ALLIANCE CREATIVE

Valencia, 6, 7 y 8 de mayo de 2016

Actas del I Congreso Internacional de Enseñanza de Inglés en Centros Educativos

Actas del I Congreso Internacional de Enseñanza de Inglés en Centros Educativos

Virginia Vinuesa y Manuel Lázaro
(Coordinadores)

CEU | *Ediciones*

Cualquier forma de reproducción, distribución, comunicación pública o transformación de esta obra solo puede ser realizada con la autorización de sus titulares, salvo excepción prevista por la ley. Diríjase a CEDRO (Centro Español de Derechos Reprográficos, www.cedro.org) si necesita fotocopiar o escanear algún fragmento de esta obra.

Actas del I Congreso Internacional de Enseñanza de Inglés en Centros Educativos

© 2016, sus autores
© 2016, de la edición, Fundación Universitaria San Pablo CEU

CEU *Ediciones*
Julián Romea 18, 28003 Madrid
Teléfono: 91 514 05 73, fax: 91 514 04 30
Correo electrónico: ceuediciones@ceu.es
www.ceuediciones.es

ISBN: 978-84-16477-51-7
Depósito legal: M-42220-2016

Maquetación: Servicios Gráficos Kenaf, S.L.

COMITÉ ORGANIZADOR

Xavier Gisbert: Dirección

Virginia Vinuesa: Programas y contenidos

Mª José Martinez de Lis: Relaciones Institucionales

Franc Corbi: Organización

Fernando Serrano: Coordinador de acreditaciones académicas

COMITÉ CIENTÍFICO

Adalid Ruiz, Pedro – Comunidad Valenciana

Aguilera Lucio-Villegas, Carmen – Comunidad de Madrid

Ball, Phil – Universidad del País Vasco

Biringer, William John – Florida State University

Casal Madinabeitia, Sonia – Universidad Pablo de Olavide

Cerezo Herrero, Enrique – Universidad CEU Cardenal Herrera

Colomar Gisbert, Salvador – Comunidad de Valencia

Cornelio, María – Hunter College (CUNY)

Cunningham, Kristina – European Commission

De Haro Figueroa, Trinidad – Ministerio de Educación

Escobar Artola, Lilly – Universidad CEU Cardenal Herrera

Fernández Fernández, Raquel – Centro Universitario Cardenal Cisneros

García Laborda, Jesus – Universidad de Alcalá

García Manzanares, Nuria – Universidad Rey Juan Carlos

García Mayo, María del Pilar – Universidad del País Vasco

García Perales, Vicent – Universidad CEU Cardenal Herrera

Genessee, Fred – McGill University, Montreal, Canadá

Gisbert da Cruz , Xavier – Comunidad de Madrid

Henderson, Rosalie – Universidad Rey Juan Carlos, Madrid

Lara Garrido, Manuel F. – BEP Network manager, Jaen

Lasagabaster, David – Universidad del País Vasco

Lorenzo Galés, Nieves – Generalidad de Cataluña

Madrid Fernández, Daniel – Universidad de Granada

Matoses Jaén, Sara – Universidad CEU Cardenal Herrera

Medgyes, Péter – Eötvös Loránd University, Budapest, Hungría

Noguera Borel, Alejandro – Fundacion Cañada Blanch y Fundacion Libertas 7

Nordlund, David – Florida State University

Palfreeman, Linda – Universidad CEU Cardenal Herrera

Palma Fernández, Gracia – Presidenta de Greta

Reyes, Charo – GRETA

Renart Ballester, Alejandra – Comunidad Valenciana

San Isidro Agrelo, Xabier Asesor – Consejería de Educación en el Reino Unido

Stobbs, Janet – Universidad CEU Cardenal Herrera

Tarrant Brown, Patricia – Bilingual Education Consultant, Valencia

Villoria Prieto, Javier – Universidad de Granada

Vinuesa Benítez, Virginia – Universidad Rey Juan Carlos, Madrid

17

Working with Corpora in TEFL

MARTA GARROTE. DEPARTMENT OF LANGUAGES AND LANGUAGE TEACHING, UNIVERSIDAD AUTÓNOMA DE MADRID

Resumen

El presente artículo tiene como finalidad la iniciación de profesores no familiarizados con la Lingüística de Corpus en el uso de ésta para la Enseñanza del Inglés como Lengua Extranjera (EILE), con el objetivo de aportar conocimientos sobre las características básicas de la disciplina, nociones esenciales sobre corpus y, especialmente, cómo aplicar los corpus en las clases de ILE. En los últimos años se ha insistido en la necesidad de usar material auténtico y recurrir a ejemplos de lenguaje real en las clases de lengua extranjera, y los corpus son una inmensa fuente de lengua en contexto real transformable en material docente, como se ejemplifica en el presente trabajo. El uso de corpus en EILE es útil no sólo para utilizar ejemplos de lenguaje real, sino también para promover la enseñanza semipresencial y basada en tareas y fomentar un aprendizaje autónomo.

Palabras clave:

Lingüística de Corpus; EILE; enseñanza basada en corpus; aprendizaje centrado en el alumno; aprendizaje autónomo.

Abstract

This paper aims to introduce the use of corpora in Teaching English as a Foreign Language (TEFL) to those teachers who are not familiar with Corpus Linguistics, with the purpose of providing insight into the key characteristics of the discipline, basic notions on what a corpus is and, more importantly, how corpora can be applied to EFL classes. In recent years the need of using authentic material and resorting to real language samples in foreign language classes has been emphasised by researchers, and corpora are a huge source of language in real context which can be transformed into teaching material, as exemplified in this paper. Working with corpora in TEFL is useful not only to handle real language examples, but also to promote task-based and blended learning, and foster students' autonomous learning.

Keywords:

Corpus Linguistics; TEFL; corpus-based teaching; learner centredness; autonomous learning.

Introduction

Over the last decades much has been written on Corpus Linguistics (CL henceforth). Corpus-based studies are numerous and many research groups worldwide devote their work to investigate this discipline (for a review see Kennedy, 2014; McEnery & Hardie, 2011; O'Keeffe & McCarthy, 2010). Probably, any researcher on areas such as Linguistics, Computer Science or Literature, among others, know about CL. However, leaving aside those who have expertise or any experience working with corpora, who does really know what a corpus is and, what is more important, how to take advantage of it for research or teaching purposes?

On many occasions, talking to colleagues who research on language, literature or education, they acknowledge that, though knowing what a corpus is and being aware of the growth of corpus-based studies and its usefulness, they lack information on how to build a corpus or, once collected, how to make use of it by automatic means, profiting from the appropriate computational tools. This tasks may seem obvious for many researchers, especially for experts (to whom this paper might be useless), but not for others. Although CL, closely bound to Computational Linguistics, involves, as any other discipline, many years of training and knowledge acquisition, there are some relatively simple tasks that may be of great help to researchers and teachers, and that can be easily performed by non-experts on CL.

Literature concerning the use of corpora for language teaching highlights the gap between research — usually carried out in university contexts— and teaching in tertiary education, as well as in other teaching settings (Campoy, Cubillo, Belles-Fortuno & Gea-Valor, 2010; Gabrielatos, 2005; McEnery & Xiao, 2011; Meunier, 2011;Römer, 2010; among others). Despite the huge amount of research on the pedagogical use of corpora and its proved success in terms of academic results (Bernardini, 2002; Keck, 2004; Römer, 2008; Smitterberg, 2004), there is still reticence by teachers to use corpora as part of their methodology, maybe due to shortage of effective training or information (not to mention insufficient technological resources or some other educational deficiencies beyond teachers' responsibility).

This paper aims to, on the one hand, be a bridge between CL and EFL teachers who, being non-experts in CL, would like to make use of it to improve their teaching practice; and, on the other, to show some easy examples of how to manage a corpus by computer means at an easy level. After a brief review of essential concepts, the focus will be placed on CL related to TEFL, and some basic methods of building and using a corpus will be showed and explained in detail. Though this work is focused on TEFL, the methods presented to exploit a corpus can be applied to other disciplines.

A brief history of CL

There are two main areas of linguistic studies: traditional prescriptive linguistics and studies of language use or descriptive linguistics (Biber, Conrad & Reppen, 1998). The basic difference is that prescriptive linguistics aims to establish a standard language and prescribe what is supposed to be the correct linguistic variety. However, descriptive linguistics observes how language is actually used by the different linguistic communities. Freeman & Freeman (2004, p. xiv) support a descriptive approach to language teaching 'because prescriptive approaches to natural phenomena like language simply don't work'. In this sense, CL meant a substantial change in linguistic research, allowing for an objective study of language based on empirical data (McEnery & Wilson, 1996; McEnery & Xiao, 2011). CL studies language use: it is the study of natural authentic language, how, when and why language is used. Corpora, or large principled collection of natural texts, represent the appropriate resource for that aim.

The path towards the last generation of corpora has evolved linked to technological advance. We can talk about Corpora B.C. (before computers) and Corpora A. C. (after computers) (Francis, 1992). Before the computer age, corpora were collected, stored and analysed manually, and one of the most common purposes was describing a particular language with the ultimate aim of teaching (McEnery & Wilson, 1996). Thorndike (1921) manually compiled the list of the 30,000 most frequent words of English (based on a four and a half million words corpus) and a lot of subsequent work on vocabulary was based on his research. Another example was the *Survey of English Usage* (Quirk, 1960), a non-electronic corpus of spoken and written English which can be considered the basis of modern corpora. Randolph Quirk and his team compiled 200 texts of 5,000 words each: 1,000,000 million words.

The corpus became electronic only in the 80s with the spread of personal computers (PCs). PCs made possible the collection, storage, processing and analysis of huge amounts of data, giving birth to Modern Corpus Linguistics, exemplified by the London-Lund Corpus (Svartvik, 1990). Today, CL is the basis of a huge number of studies in language description, linguistic theory, language teaching and learning, translation, natural language processing, etc.

But, what is CL? There is some disagreement between those who believe that CL is another branch of Linguistics, at the same level of Sociolinguistics or Descriptive Linguistics, to cite some; and those who think CL is just a methodology, a tool to analyse texts. Actually, not only does it make an incredibly useful tool, but also has its own theoretical basis as well (McEnery & Wilson, 1996). In any case, CL is the linguistic study on the basis of text corpora, 'which focuses upon a set of procedures, or methods, for studying language' (McEnery & Hardie, 2011, p. 1).

What is a corpus?

A corpus is not just any collection of texts, but a *principled* collection of *natural* texts (Biber et al., 1998) that can be used for linguistic analysis. *Natural* means that texts must be authentic, not specifically written or read to create a spoken or written corpus. Therefore, a written corpus must be made up of real texts from newspapers, academic writings, literary texts, and so on; and a spoken corpus must be built upon real speech

that has to be recorded and transcribed. There are five essential requirements a good corpus must fulfil (Biber et al., 1998; Kennedy, 2014; McEnery & Wilson, 1996):

- Representativeness: the corpus must be large enough to represent the type of texts or the variety of language it stands for. The suitable size may vary according to the corpus type —for spoken general corpora, anything over a million words is considered to be large; for general written corpora, anything below five million is quite small (O'Keeffe, McCarthy & Carter, 2007).
- Electronic format: it must be stored in a computer.
- Sampling principles: the design must take into account the type of corpus to be built and clearly establish the genre, quantity (to be representative) and features of the texts to be collected.
- Internal structure: the architecture of a corpus must be well-defined, as a whole, as well as taking into account every possible subcorpora or the organization of texts. For instance, in spoken corpora, at the beginning of each text there should be a header with information about the participants' sociolinguistic features and the characteristics of the communicative situation.
- Available information about the corpus characteristics and data: the user must have access to an introductory document in which information about the size, type of texts, number of participants, etc. can be consulted.

Corpora can be general (also called reference), which get together representative samples of a language, or specialized, collecting samples of a particular field of expertise to study a specific phenomenon (Römer, 2010). Corpora can also be multilingual, usually built for Comparative Linguistics studies or Translation studies, or one-language corpora. They can also be spoken or written. Large corpora usually have a spoken subcorpus, but it is always smaller than the written one (for example, in the British National Corpus, which is made up of one hundred million words, only a seventeen percent is spoken language, and the rest, eighty two percent, is written). This is because corpora of spoken language are much more time-consuming to assemble. It takes considerably longer to build because speech has to be recorded, transcribed and possibly coded for some of its non-verbal features. And finally, a corpus can be just orthographic or it can also be annotated, that is, enriched with information as Part of Speech, syntactic and semantic information, etc.; if it is a spoken corpus, it may also contain information on prosody features, for example.

The following list present some of the most prominent corpora.

- General:
  - BNC (British National Corpus): 100 million words. Different types of spoken and written texts (http://corpus.byu.edu/bnc/).
  - COCA (The Corpus of Contemporary American English): 450 million words. Spoken and written text types (http://corpus.byu.edu/coca/).
- Specialized
  - COLT (Corpus of London Teenage Language): 500,000 words. Spontaneous spoken language (http://clu.uni.no/icame/colt/).
  - ICLE (International Corpus of Learner English): 2.5 million words. Argumentative essays written by learners with different L1s (https://www.uclouvain.be/en-cecl-icle.html).
  - CHILDES (Child Language Data Exchange System): first language acquisition corpora. Over 20 languages (http://childes.psy.cmu.edu/).

The British National Corpus (BNC) and its American counterpart (COCA) are two examples of general corpus, as they are representative of the different texts (spoken and written) of a language. However, COLT, ICLE and CHILDES are examples of specialized corpora. All of them are freely consulted through an interface especially developed to that aim or through direct download.

Corpus uses: the pedagogical corpus

Despite some detractors to CL, its relevant contribution to the study of language is currently undeniable. Corpus-based studies abound in language research and analysis, in lexicography (creation of dictionaries whose vocabulary is based on the word list of one or more corpora), in the elaboration of grammars based on real use, in translation (multilingual parallel corpora) and, among others, in language pedagogy, especially in second

language teaching (McEnery & Xiao, 2011). There are two major areas of CL research on Foreign Language Teaching (FLT):

- New descriptions of language (common learners' mistakes, influence of L1 on the target language, inter-language, etc.)
- Production of language teaching materials (vocabulary lists based on real use, grammars, idiomatic expressions, language use).

But strictly regarding teaching, scholars mention two methods of using corpora: the indirect and the direct methods (Gabrielatos, 2005; McEnery & Xiao, 2011; Römer, 2008, 2010). The former refers basically to materials development and the latter to what McEnery and Xiao (2011, p. 365) call 'teaching about, teaching to exploit, and exploiting to teach', and, ultimately, to the building of teaching-oriented corpora (Language for Specific Purposes corpora or learner corpora). As the authors explain, 'teaching about' refers to teaching CL as an academic subject; and 'exploiting to teach' is related to teaching linguistics by using a corpus-based approach. It is 'teaching to exploit' the concept in which we, as EFL teachers, are interested, as it involves providing the language students with the necessary knowledge to work with corpora, so they become researchers (Johns, 1991). This idea leads to the term 'discovery learning', encouraged firstly by Johns (1991) and his Data-Driven Learning (DDL) method, by means of which students explore language themselves. Depending on the students' level, DDL activities can be teacher-directed or learner-led, but always learner-centred. The benefits of DDL are numerous: there is a change from the traditional 'three P's' (presentation-practice-production) to the 'three I's' (illustration-interaction-induction) (McEnery & Xiao, 2011); students are in charge of their own learning, increasing their motivation (Keck, 2004); DDL puts 'the learner (instead of the teacher) at centre stage' (Römer, 2010, p. 20); the students analyze authentic language examples (Braum, 2006); the use of corpora 'helps them [students] develop longer-term cognitive skills (such as a greater awareness for lexico-grammatical aspects), and promotes independent learning' (Meunier, 2011, p. 463). According to Römer (2010, p. 26) DDL leads to 'learner motivation, serendipity, communicative competence, language awareness raising and learner autonomous learning'.

DDL activities can be carried out on native speakers corpora (either general or specialized) from a top-down approach, or on learner corpora —'systematic computerised collections of the language produced by language learners' (Römer, 2008, p. 117)—, from a bottom-up approach, by analyzing common mistakes and even comparing learners' real productions to native speakers' ones. The website of the Université Catholique de Louvain provides a complete list of learner corpora with relevant information not only about the corpora language/s, but also regarding their size, texts type or availability (https://www.uclouvain.be/en-cecl-lcworld.html).

There is not enough room here to develop the productive theoretical research on the relation between teaching and CL. Therefore, we go on the aim of this paper, which is giving specific guidance for teachers to apply CL in their teaching practice, indirectly and directly, with concrete examples explained in detail. For further information, we encourage the reader to delve into the reference section.

Corpus tools

Obviously, the main advantage of modern corpora is that texts can be handle by means of computational tools, which facilitate work mainly in two senses: firstly, time-consuming manual tasks can be performed by a computer in seconds; secondly, a computer guarantee the reliability of results in terms of quantitative data (minimizing the human margin of error). To manage corpora we count on different software: *annotation tools*, which add information at different levels of the text (part of speech, syntax, semantics); and *analysis tools*, which are query systems to look for words and their distribution in a text (frequency lists, keywords, collocations). The upsurge of this kind of software is currently overwhelming. Therefore, only a small part of them will be mentioned.

Annotation tools

We count on a huge amount of annotation tools to tag different linguistic levels, from the most elementary, a PoS (Part-of-Speech) tagger, to the most challenging semantic or pragmatic annotation. Nowadays, PoS taggers are robust programs with a low margin of error. Some well-known examples are CLAWS (Garside, 1987) or the Stanford Part-of-Speech Tagger (Toutanova, Klein, Manning & Singer, 2003). The main disadvantage of these type or resources is that they tend not to be user-friendly, and once the software is download, their installation

and use is not intuitive. Furthermore, some of them are not free and licence fees may be expensive. Anyway, there are two good options for beginners: the free CLAWS PoS tagger, with a user-friendly interface where the user just copies a text and automatically obtains a grammatical tagging of every word (http://ucrel.lancs.ac.uk/claws/); and SMILE Text Analyzer, with the same operating mode (https://smile-pos.appspot.com/).

Analysis tools

Analysis tools are corpus query systems which allow for looking at how words behave in texts. The basic tasks these kinds of tools provide are word lists (the vocabulary included in a text, ordered by frequency or alphabetically), keywords (words whose frequency is unusually high in a given corpus.), collocations (sequences of words that co-occur more often) or concordance lines (a particular word in context). Two of the most outstanding are *WordSmith Tools* (Scott, 1996) and *Sketch Engine* (Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). However, a perfect option for beginners is *AntConc* (Anthony, 2014), a free analysis tool characterized by its simplicity and robustness[1].

Closely related to DDL tasks, linguistic analysis software is the most useful tool of both explained here for TEFL aims, as we will see in the next section.

Introducing corpora in the TEFL classroom

To sum up the information presented so far, the use of corpora by TEFL teachers is synthesized in Figure 1.

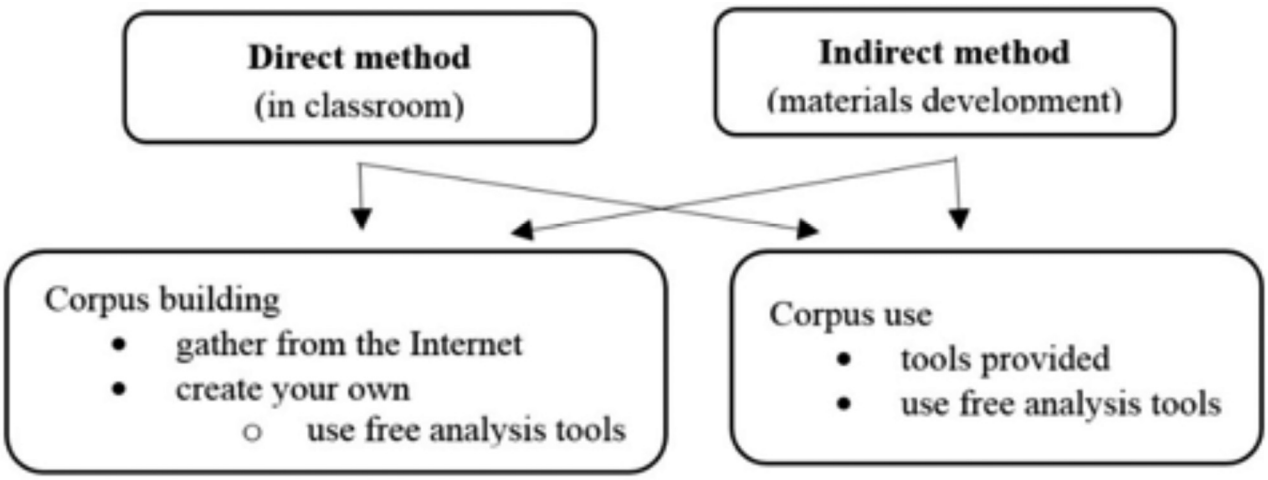

Figure 1. Use of corpora by TEFL teachers.

TEFL teachers can either use corpora to create their own teaching material, selecting what to teach, or to directly develop tasks in the classroom, deciding how to teach. For both methods, either an existing corpus can be used or teachers can develop their own, according to their needs. Next, examples of all possibilities will be presented and explained.

Direct method-corpus use

Concerning using an existing corpus to use it directly with the students, there are two possibilities: making use of the analysis tool provided in the corpus website; or downloading the corpus (if possible) and using any free-downloadable analysis tool to handle the corpus. Let us have a look at two corpora, both providing analysis tools.

The first one is the Michigan Corpus of Academic Spoken English (MICASE) (Simpson, Briggs, Ovens & Swales, 2002). The project's home webpage has the following look:

[1] http://www.laurenceanthony.net/software/antconc/

Figure 2. MICASE home webpage

By clicking on the link Search MICASE, an interface will be opened (Figure 3). In it, speakers' and texts' features can be selected, if necessary, to constrain the search.

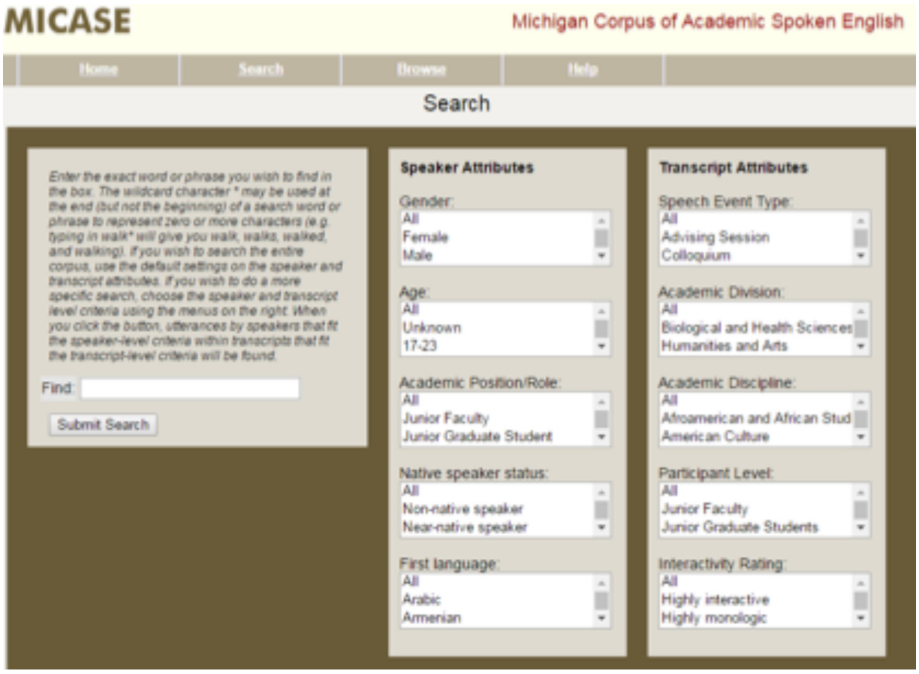

Figure 3. Search MICASE

Students can be asked to search for a specific word to analyse its linguistic distribution. Let us write the verb 'make', and the analysis tool will retrieve all the examples of this word in context —what is called *concordance*— found in the corpus (2,215 occurrences). This is a DDL task for students to explore the uses of the word 'make'.

Figure 4. Uses of 'make' in MICASE

The second corpus that TEFL teachers can explore with their students is Netspeak (Riehmann, Gruendl, Froehlich, Potthast, Trenkmann & Stein, 2011). Actually, it is not a corpus in itself, but a web search engine. Therefore, the corpus is the total number of texts from the Internet. Netspeak aims to assist the writer and uses the World Wide Web to retrieve examples of a certain query, allowing for the use of regular expressions[2] explained in their home webpage. A possible activity is that students look for the pattern *such + any word + that* and *so + any word + that* (Figure 5) and try to induce the norm or grammatical rule (*any word* is expressed by the regular expression '…').

Figure 5. Netspeak search

Direct method-corpus building

Creating a corpus is not a trivial project, as it was explained above. However, it is possible to build *ad hoc* rudimentary corpora for teaching purposes. Students can be asked to assemble a corpus from the Internet, searching for texts on a particular topic. They must copy the texts and paste them into a text editor, saving them as plain text (.txt)[3]. Thus, a small corpus on, for example, *Traveling*, is collected.

The next step is downloading AntConc and running it in the computers they are working with (http://www.antlab.sci.waseda.ac.jp/software.html).

[2] In computer science, a regular expression is a sequence of characters that forms a search pattern. In that sequence wildcard characters can be added to the word or phrase looked for.

[3] Most annotation and analysis tools just accept plain texts.

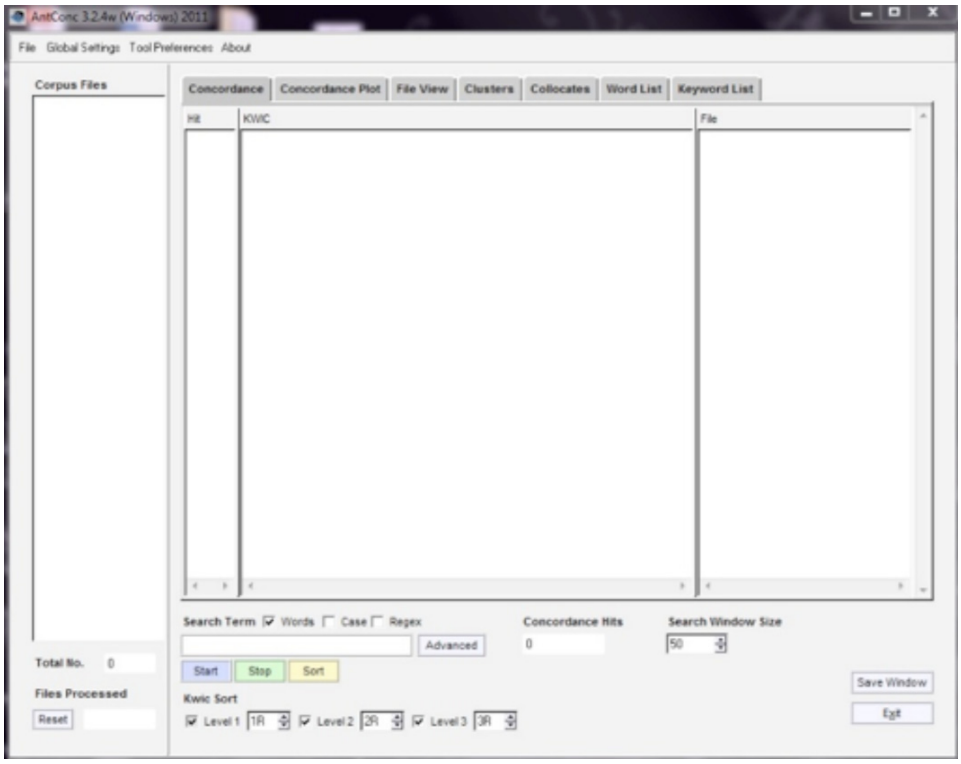

Figure 6. AntConc interface

In the AntConc interface (Figure 6) students must select 'File' and load their corpus. Then, they just have to click on the 'Word list' tab and press the 'start' button to obtain a list of vocabulary and the most frequent words on the selected topic. They can also use the tab 'Collocates' to look for sequences of words that co-occur more often in that kind of texts.

Indirect method-corpus use

Corpora are also a very useful and reliable resource to develop teaching material. For the following example the BNC is used to create lexico-grammatical profiles, useful to promote independent learning. Through the webpage http://corpus.byu.edu/bnc/ the BNC can be consulted by means of the following interface (Figure 7):

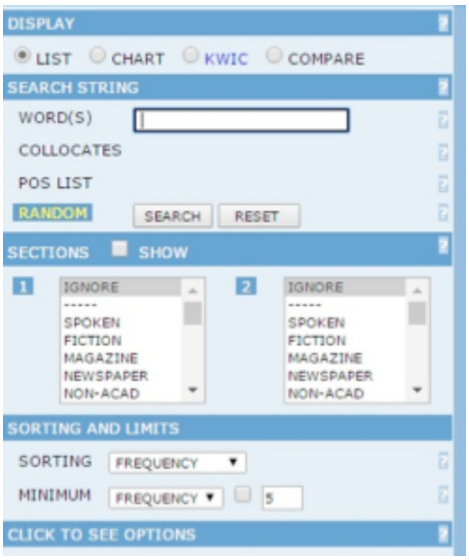

Figure 7. BNC search

The searching interface offers multiple possibilities, but for the present example the option KWIC (keyword in context) must be ticked; texts genre ('Spoken', 'Fiction', Newspaper', etc.) can be also constrained (for Figure 8 results the selected genre was 'Academic'). Let us type the word 'test' to see an example.

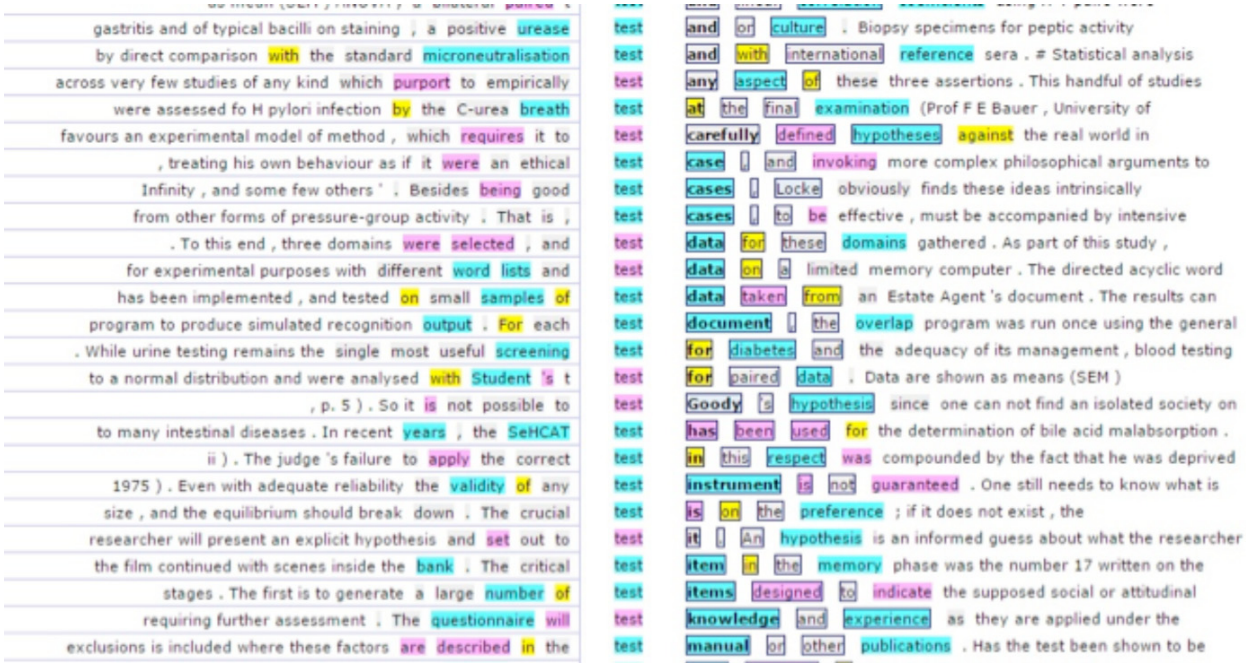

Figure 8. Results of the word 'test' from the BNC

Retrieved results exceeded 3,000 occurrences, a few of them presented in Figure 8. The searched word is centrally placed surrounded by co-occurring left and right words. Furthermore, the word itself and co-occurring words are PoS tagged (nouns in blue, verbs in pink, prepositions in yellow, and so on).

This example of use retrieved from the BNC, together with a table as that presented below (Table 1), can be given to students as a worksheet. They just have to fill the table gaps with the information required by analysing the corpus results and, while doing it, they will learn vocabulary in a more meaningful way than just memorizing word lists or doing training exercises or drills. In Gabrielatos (2005, p. 24) words, 'in consulting a dictionary or grammar learners are given fish; by actively engaging in pattern recognition they learn how to fish'.

WORD	test
PoS	(category)
Collocates	(which words co-occur most frequently?)
Chunks/Idioms	(recurrent chunks? Idioms?)
Syntactic restrictions	(prepositions, clause-position)
Semantic restrictions	(applied to humans, never used with an intensifier)
Other relevant features	

Table 1. Lexico-grammatical word profile

Indirect method- corpus building

The last example describes a method of creating teaching material by means of building a corpus. Collecting student's production examples to build a corpus is quite interesting and useful, not only from a procedure viewpoint, but also for assessment issues. Though a spoken corpus would involve more difficulty and working hours (recording students and transcribing their speech), a written one can be easily gathered by asking students to submit their writing tasks by email. These writings can be structured according to topic, student's level, type of activity, etc. Once the corpus is ready, an analysis tool as AntConc can be used to look for collocations, frequent words or even keywords, which will give the teacher a wide view of his/her students' knowledge and further needs. As an example, the *Gachon Learner Corpus* (Carlstrom & Price, 2014) was downloaded and used to retrieve information using AntConc. This corpus is made up of writings from 2,500 Korean, Chinese and Spanish students and more than 25,000 texts. To exemplify the procedure, just 100 texts were selected and loaded in AntConc (Figure 9).

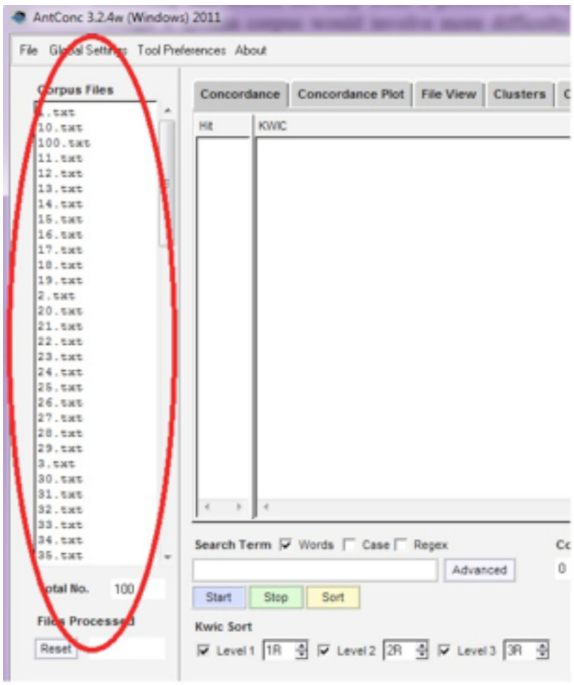

Figure 9. Selected files from the Gachon Learner Corpus

Selecting the tab 'Concordance', we can look for our students' typical mistakes, i.e. the use of the *verba dicendi* 'say', 'tell' and 'ask', so common in reported speech and common source of confusion for basic and intermediate students. Curiously, the search result for 'tell' (Figure 10) showed a non-targeted kind of error. This is a very useful procedure to detect students' needs and even to generate a list of typical EFL learners' mistakes on which teachers must focus.

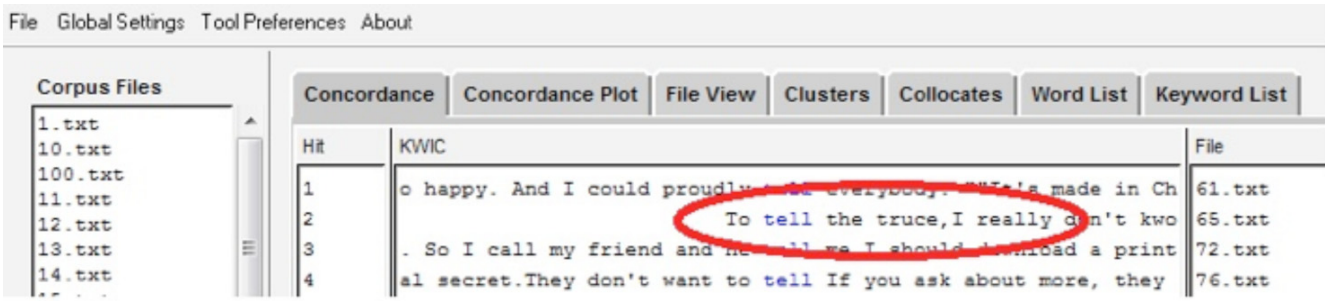

Figure 10. AntConc seach

Your own students' corpus can be used to develop teaching material by, for example, elaborating a worksheet with their mistaken productions and ask them to try to find the mistakes and correct them. Using AntConc it is as easy as saving the search results (Figure 11). As Römer states (2008, p. 121) 'this focus on familiar texts (i.e. on texts the learners themselves have produced) ensures motivation'.

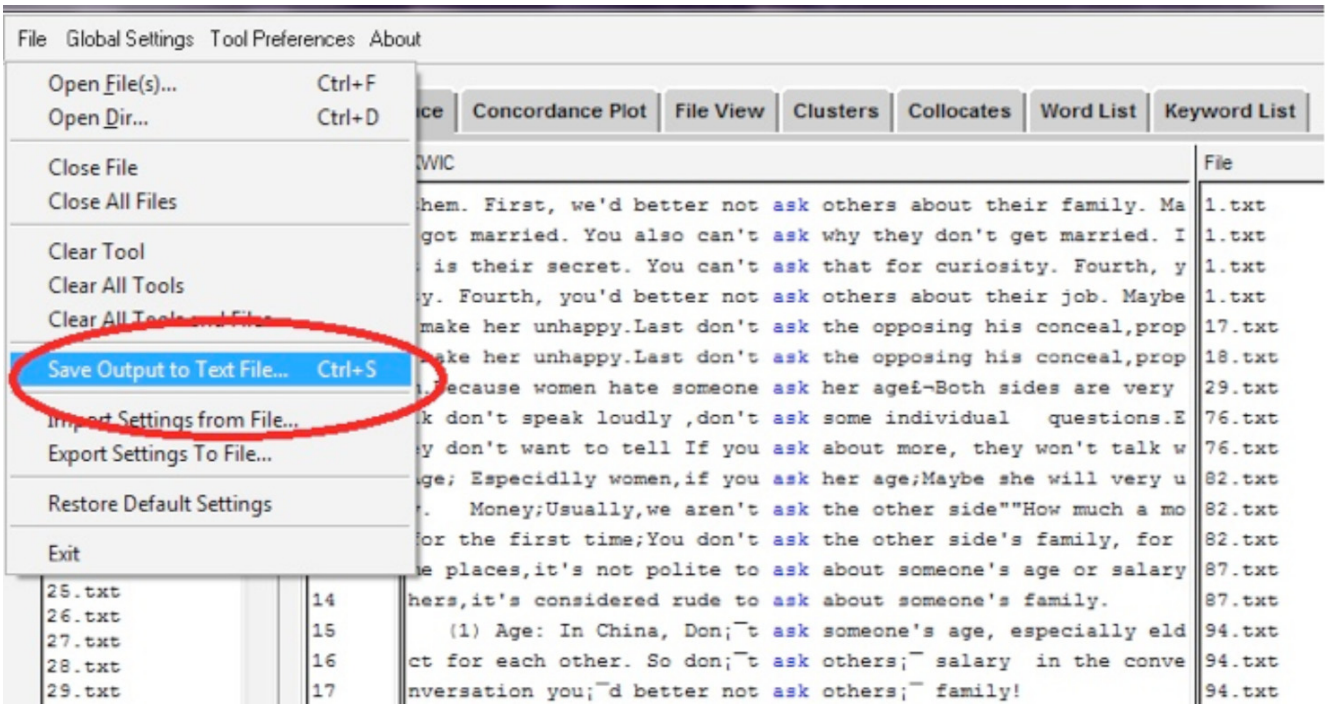

Figure 11. Saving AntConc results

Final considerations

So far, some examples of how EFL teachers can profit from CL have been presented. These mean just a tiny representation of the immense possibilities CL can offer to TEFL, as increasing literature on the topic evidences (Aston, Burnard, McEnery, Granger, Hung, Petch-Tyson & Marko, 2004; Burnard & McEnery, 2000; Ghadessy, Henry & Roseberry, 2001; Granger, Hung & Petch-Tyson, 2002; Sinclair, 2004). The usefulness of CL for TEFL is supported by many scholars, who also complain about the scarce use of corpora in real teaching contexts (Keck, 2004; Römer, 2008; among others). It is clear that there is a gap between CL and teaching, a missing piece. Probably, one of the main reasons is the lack of training addressed to language teachers (Gabrielatos, 2005). Römer (2010) also establishes the future tasks in applied corpus linguistics, namely, concentrating on learners' and teachers' needs and promoting direct and indirect uses of corpora in language teaching. She also highlights the difference between 'computer-savvy' and 'techno-phobic' learners (terms which can be applied to teachers as well) and the need to work to overcome this issue.

The limitations of the present paper are obvious, partly due to lack of space, partly due to its purpose to be a light introduction for non-experts. A great deal of relevant contents and examples have been left out. Therefore, for those interested in the subject, I would recommend the work of Gabrielatos (2005) or O'Keeffe el al. (2007) for interesting examples on how to apply CL to TEFL; or Cobb's (2016) *Compleat Lexical Tutor* (http://www.lextutor.ca/) and Davies' (2016) YouTube channel (https://www.youtube.com/user/CorpusProf) for ready-made DDL exercises. Also, Anthony's (2014) webpage provide various analysis tools, apart from AntConc, together with clarifying tutorials. Finally, regarding learner corpora, the Written Corpus of Learner English (WRICLE) is a corpus written in English by Spanish learners built under the WOSLAC Project, developed by researchers at Universidad Autónoma de Madrid, and freely available at http://www.uam.es/proyectosinv/woslac/Wricle/.

References

Anthony, L. (2014). *AntConc (Version 3.4.3)* [Computer Software]. Tokyo, Japan: Waseda University. Available at http://www.laurenceanthony.net/

Aston, G., Burnard, L., McEnery, T., Granger, S., Hung, J., Petch-Tyson, S. & Marko, G. (2004). Corpus linguistics and language teaching research: bridging the gap.

Bernardini, S. (2002). Exploring new directions for discovery learning. *Language and Computers*, 42(1), 165-182.

Biber, D., Conrad, S. & Reppen, R. (1998). *Corpus linguistics. Investigating language structure and use.* Cambridge: Cambridge University Press.

Braun, S. (2006). ELISA–a pedagogically enriched corpus for language learning purposes. In S: Braun, K. Kohn & J. Mukherjee (Eds.), *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods* (pp. 25-47). Frankfurt/M: Lang.

Burnard, L., & McEnery, T. (2000). Rethinking language pedagogy from a corpus perspective. In *Third International Conference on Teaching and Language Corpora*. Frankfurt a. M. ua.

Campoy, M. C., Cubillo, M. C. C., Belles-Fortuno, B., & Gea-Valor, M. L. (Eds.). (2010). *Corpus-based approaches to English language teaching*. A&C Black.

Carlstrom, B. & Price, N. (2014). *The Gachon Learner Corpus:* Available at http://koreanlearnercorpusblog.blogspot.com.es/p/corpus.html.

Francis, W. N. (1992). Language corpora BC. In *Directions in Linguistics: Proceedings of Nobel Symposium* (Vol. 82, pp. 17-32).

Freeman, D. E. & Freeman, Y. S. (2004). *Essential Linguistics: What You Need to Know to Teach Reading, ESL, Spelling, Phonics, and Grammar*. Portsmouth: Heinemann.

Gabrielatos, C. (2005). Corpora and language teaching: Just a fling, or wedding bells? *TESL-EJ*, 8(4), 1-37.

Garside, R. (1987). The CLAWS word-tagging system. In R. Garside, G. Leech & G. Sampson (Eds.), *The computational analysis of English: a corpus-based approach*. London: Longman.

Ghadessy, M., Henry, A., & Roseberry, R. L. (Eds.). (2001). *Small corpus studies and ELT: theory and practice* (Vol. 5). John Benjamins Publishing.

Granger, S., Hung, J., & Petch-Tyson, S. (Eds.). (2002). *Computer learner corpora, second language acquisition, and foreign language teaching* (Vol. 6). John Benjamins Publishing.

Johns, T. (1991) 'Should you be persuaded: two samples of data-driven learning materials'. In T. Johns & P. King (Eds.), *Classroom Concordancing ELR Journal 4*. University of Birmingham.

Keck, C. M. (2004). Book Review: Corpus linguistics and language teaching research: bridging the gap. *Language Teaching Research January*, 8, 83-109

Kennedy, G. (2014). *An introduction to corpus linguistics*. Routledge.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), 7-36.

McEnery, T. and Wilson, A. (1996) *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

McEnery, T., & Xiao, R. (2011). What corpora can offer in language teaching and learning. In E.Hinkel (ed.), *Handbook of research in second language teaching and learning*, Vol. 2, 364-380.

Meunier, F. (2011). Corpus linguistics and second/foreign language learning: exploring multiple paths. *Revista Brasileira de Linguística Aplicada*, 11(2), 459-477.

O'Keeffe, A., McCarthy, M. and Carter, R. (2007) *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge University Press.

O'Keeffe, A., & McCarthy, M. (Eds.). (2010). *The Routledge handbook of corpus linguistics*. Routledge.

Quirk, R. (1960). The survey of English usage. *Transactions of the Philological Society*, 70-87.

Riehmann, P., Gruendl, H., Froehlich, B., Potthast, M., Trenkmann, M., & Stein, B. (2011). The NETSPEAK WORDGRAPH: Visualizing keywords in context. In *Visualization Symposium (PacificVis)*, 2011 IEEE Pacific, 123-130).

Römer, U. (2008). Corpora and language teaching. *Corpus linguistics. An international handbook*, 1, 112-131.

Römer, U. (2010). Using general and specialized corpora in English language teaching: Past, present and future. *Corpus-based approaches to English language teaching*, 18-38.

Scott, M. (1996). *WordSmith tools*. Oxford: Oxford University Press.

Simpson, R. C., Briggs, S. L., Ovens, J., & Swales, J. M. (2002) *The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI: The Regents of the University of Michigan.

Sinclair, J. M. (Ed.). (2004). *How to use corpora in language teaching*. John Benjamins Publishing.

SMILE Text Analyzer (2016). Available at: https://smile-pos.appspot.com/ (accessed 10 April 2016).

Smitterberg, E. (2004). Review of: Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching. *ICAME Journal*, 28, 114-119.

Svartvik, J. (Ed.). (1990). *The London-Lund corpus of spoken English: Description and research*. Lund University Press.

Thorndike, E. L. (1921). *A teacher's word book*. New York: Columbia Teachers College.

Toutanova, K., Klein, D., Manning, C. & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL*, 252-259.