Krajka, Jarosław (2007) Corpora and Language Teachers: From Ready-Made to Teacher-Made Collections. *CORELL: Computer Resources for Language Learning* 1, 36-55.

Corpora and Language Teachers: From Ready-Made to Teacher-Made Collections

Jarosław Krajka
Warsaw School of Social Psychology, Poland

Abstract. The prevalence of computers, increased opportunities of Internet access, availability of large amounts of target language data, all of these call for language teachers' greater interest in the active use of corpora both for the classroom (in materials development) and in the classroom (for learner discovery tasks). Contrary to the pre-Internet era, when corpus consultation procedures were largely restricted to linguists and lexicographers due to technological, financial and logistical considerations, the language teachers of the Web 2.0 age will find it much easier to access, compile and consult corpora for language teaching.

The aim of the present paper is to provide an overview of available corpus resources, proceeding from ready-made corpora to be consulted on the Internet by teachers of English, French and German, through the approach and tools for using the Web as a Corpus (WaC), finally proceeding to the compilation of custom-made ("ad-hoc") corpus collections to suit the needs of particular teaching contexts. The paper concludes with a comparative review of a few widely available concordancers to be used with teacher-made corpora.

Keywords. Corpus linguistics, concordancing, ad-hoc corpora, Web as a Corpus.

1. Introduction

Corpora and concordancing have been widely used in ELT by materials writers and lexicographers, also to assist teachers in making informed choices about curriculum development, vocabulary selection and lexical testing. However, these tools were not of wide use by teachers, mainly due to lack of proper information and training, largely limited access, as well as lack of full relevance for some contexts (e.g., LSP). On the other hand, promoting active learning through exposure to "raw" linguistic data produced by concordancers helps to increase the inductive reasoning skills, enabling students to formulate rules on their own and retain them for a longer period of time.

With a computer being a standard tool of a contemporary language teacher, together with greatly facilitated Internet access and much higher bandwidth, the reflection on the incorporation of concordancing procedures by language teachers in materials development or vocabulary selection seems to be of prime importance. The possible impact of hands-on concordancing on teacher autonomy (and, in effect, learner autonomy), resulting in much greater awareness of the teaching content and judicious use of coursebook materials, is another factor that calls for wider implementation of concordancing, on various planes and in different respects, in actual teacher training.

The aim of the present paper is not to provide a comprehensive discussion of all the major issues involved in corpus linguistics or Data-Driven Learning (for this, see McEnery & Wilson 1996; Biber *et al.* 1998; Kennedy 1998; or Partington 1998; to mention just a few). Instead, we would like to provide an overview of widely accessible corpora resources, categorising them into the three groups: ready-made corpora available on the Web, corpora made of online materials and teacher-made corpora. We will start with the brief definition of a corpus, followed by a discussion of the opportunities and drawbacks of corpus-based language learning, finally, proceeding to the overview of Web-based corpora resources and concordancing tools.

2. Opportunities and drawbacks of in-class corpus consultation procedures

In introduction, some space should be devoted to the definition of a corpus. Crystal (1991) defines it as "a collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting-point of linguistic description or as a means of verifying hypotheses about a language". In a similar vein, Sinclair (1991) adds that corpora are made of naturally occurring language, while Krishnamurthy (2001) points out the genuine communicative situations that are recorded without any editing to create corpus contexts. McEnery and Wilson (1996) enumerate four crucial characteristics of a corpus, being sampling and representativeness, finite and fixed size, machine-readability and standard reference. Four other criteria indispensable for a body of texts to be met to be labelled a corpus are, according to Sinclair (1996), considerable quantity, authentic quality, plain text simplicity of data storage and documentation of full details about the constituents of a component (annotation) kept separately from the document itself. Kilgariff and Grefenstette (2003) opt for a broader definition, terming a corpus "a collection of texts when considered as an object of language or literary study", thus avoiding strict criteria of McEnery and Wilson (1996), and, consequently, allowing regarding the Web as a corpus. Atkins *et al.* (1992) make an important distinction into an archive (an electronic collection of texts, not connected with one another), an electronic text library (ETL - a standardised collection of texts in an electronic format with some assumptions on the content, but without rigorous criteria of selection), and a corpus, a subtype of ETL, compiled according to clearly specified criteria for a particular purpose. Thus, what makes a collection a corpus is a clearly defined purpose that one has in mind when gathering samples of language.

Concordancing procedures, or formulating queries to retrieve occurrences of linguistic data from a corpus, can constitute a significant teaching aid both in teacher preparation for classes, in their self-development as well as in building students' language awareness ('direct' corpus use - Leech 1997). Some arguments for implementing the tool in teaching and learning might be the following:

- concordancing interjects authenticity (of text, purpose, and activity) into the learning process, as students assume control of that process and their competence is built by gaining access to the facts of linguistic performance (Johns 1988);
- noticing a word in several contexts extends the knowledge of that word and promotes successful learning (Cobb 1998);
- as a corpus is built from many texts, it displays words in many more situations than just the most prototypical ones that coursebook authors may come up with (Cobb 1998);
- the diversity of a language can never be fully presented in a dictionary, and only few dictionaries provide sufficient amount of information about a word's grammar or its collocations, while it takes a concordancer only seconds to search a corpus and give more language data (Thomas 2003);
- language students, teachers, translators and people writing in a foreign language find as invaluable help the opportunity to get access to data for checking one's intuitions on the fly (Thomas 2003);
- knowledge encoded from data by learners themselves will be more flexible, transferable, and useful than knowledge encoded by experts and transmitted to them by an instructor (Cobb 1999);
- corpus-based procedures create conditions for internalizing certain abstract grammatical concepts, such as part-of-speech or part-of-sentence distinctions (Godwin-Jones 2001);

- sociolinguistic competence is addressed by drawing attention to the issue of register through analysis of actual language use (Krieger 2003).

At the same time, one needs to be aware of at least some difficulties and obstacles of the process:

- lexical information may be vast and confusing to learners, and even though words appear in rich contexts, many of the words in the contexts are certainly unknown (Cobb 1998);
- the contexts are rich, varied and plentiful but they are also short, incomplete, and do not form a coherent whole (Cobb 1998);
- concordancers are not tools to be used by computer novices without any instruction nor preparation, and in order to formulate more efficient searches, one has to undergo proper training (Stevens 1995);
- inherent limitations in the database are rarely intuitively understood by learners, who will treat a corpus as another dictionary (Stevens 1995);
- as it is difficult for language learners to independently formulate queries to observe subtle patterns in language, the role of the teacher as a facilitator is indispensable (Stevens 1995);
- not all learners may have equally positive attitudes towards inductive discovery learning (Krieger 2003);
- careless overreliance on corpora may give a false impression of language, as corpora rarely, if at all, feature all the samples of language that are most preferable for classroom teaching (Dellar nd.).

Additionally, Godwin-Jones (2001) points out to some technical considerations that may make teacher-directed concordancing a dubious experience:

- the form in which most corpora are stored (most annotated in SGML and housed on large Unix servers, not practical to store such large amounts of data locally);
- the access provided remotely, which may present performance issues;
- the proliferation of different formats for accessing corpora and the bewildering array of tools available;
- often poorly designed interfaces of Web tools.

Thus, the awareness of the strengths of concordancing as presented above makes it possible to structure successful learning activities with significant potential. On the other hand, the reflection on the problems and technical shortcomings should lead to structuring the activities to prevent those imperfections.

3. Concordancing-based activities for EFL

The implementation of any teaching tool or procedure, not only corpora and concordancers, in a particular teaching context needs to take into account psychosocial considerations (learners' and teacher's perceptions of the tool and the process), logistical considerations (the availability of computer lab with a display device for concordancing sessions, the copying provisions for the preparation of corpora-based teaching materials) and administrative considerations (the accordance of a particular corpus or the discovery learning

approach in general with the overall training philosophy, ultimate teacher objectives and the like). Having determined these, one could reflect on the mode of use, which is where Leech (1997) distinguishes between the direct and indirect use of corpora in teaching: teaching about corpora, teaching the use of corpora and using corpora to teach are said to represent the direct use of corpora, while reference publishing, materials development and language testing are indirect applications (Leech 1997). Barlow (2002) defines and further describes three specific areas for teacher-directed corpora use:

1. syllabus design: conducting an analysis of a corpus relevant to a purpose of a given class to obtain frequency and register information to assist in course planning choices;
2. materials development: creating exercises based on real examples which provide students with the opportunity to discover features of language use;
3. classroom activities: hands-on student-conducted language analyses to elicit learner-made discoveries about language use.

When reflecting on how the type of corpus determines the use and the type of activities that are to be implemented, Gabrielatos (2005) advises teachers to develop the awareness of the specificity of the type of information included in the corpus. This correlation is summarised in the figure below.

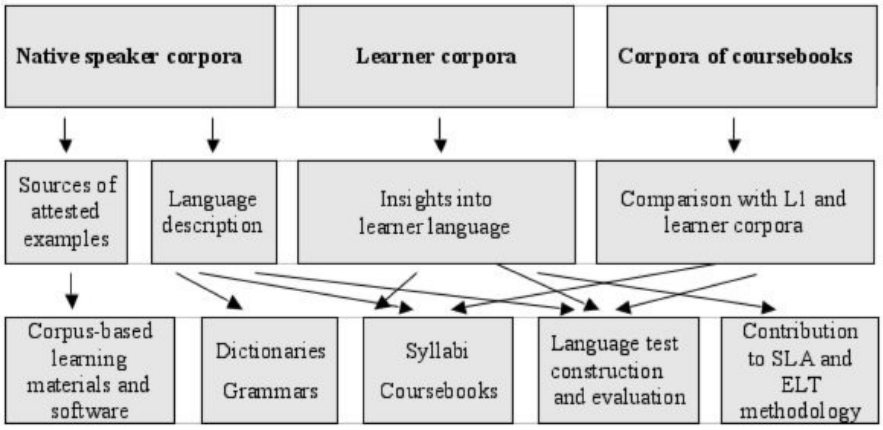

Figure 1. Uses of corpora (after Gabrielatos, 2005).

The direct use of corpora involves the teacher constructing classroom tasks or self-study learning activities, which is described in the figure above as corpus-based learning materials, where the native speaker corpora serve as a source of attested examples demonstrating real and authentic language use.

The reflection on which areas of language instruction can be subjected to corpus-based learning procedures, either with the use of ready-made or customised corpora, can most probably be interminable, leading to the formulation of a number of training activities and approaches pertinent to most aspects of the foreign language learning process. However, there are some areas where such concordance-based tasks can be of most immediate use, which can be enumerated as follows:

1. Grammar:
   - presenting new language points (Hasselgård 2001), assisting the induction of grammatical rules with concordance output on contrastive issues (e.g., the use of 'for' and 'since');

- matching expressions with speakers coming from different geographical, social, professional backgrounds;
- constructing true/false, right/wrong or error correction tasks to be verified with the use of concordance queries (see, for instance, Lextutor's "Check grammar against corpus data" activity, http://www.lextutor.ca/grammar_tester/);
- researching certain grammatical forms which may cause problems with use in order to demonstrate a greater range of applications in a fuller context (St. John 2001);
- investigating how particular language features are used, by examining their concordances and comparing frequencies (Aston 1996);
- contrasting particular constructions, demonstrating their use in various contexts (Fletcher 2001);
- practising using particular language items by interpreting citations (Aston 1996);
- verifying the prescriptive usage from the coursebook with the more contemporary, also geographically restricted, use (Fletcher 2001);
- using pre-selected concordances based on a corpus of legal texts to execute remedial grammar work (Weber 2001).

2. Vocabulary:
   - learners becoming lexicographers using computer technology to build their own dictionaries (Cobb 1998);
   - implementing vocabulary-based concordance sheets, classroom projects and task-based learning-filling gaps (Tribble & Jones, 1990; Chen 2004);
   - organising corpus-based enquiries to investigate the differences between words commonly confused or misused (Krishnamurthy 2001).

3. Reading comprehension:
   - assisting inferring new words from the text with additional, carefully selected examples (Cobb 1999);
   - making predictions about the content of the text and activating schemata in the pre-reading stage by concordancing selected words to observe the most prototypical contexts of usage.

4. Writing:
   - improving the awareness of register by using a concordancer as a look-up tool to check the register of particular words to be potentially used (or misused) in a writing piece;
   - organising learner-made comparisons of their personal corpora with publicly available collections to enhance language awareness and improve writing skills (Krishnamurthy 2001);
   - using concordances to explore possible correlations between generic structures identified and particular lexical items, to later write mini-essays incorporating the structures generated previously (Weber 2001).

4. Online corpora for EFL – from ready-made resources to teacher-made materials

The availability of ready-made corpora for widespread and unlimited pedagogical use by LSP teachers has largely increased recently, together with the popularisation of the Internet and open source software movement (Tribble 1997). The proliferation of corpora resources

demands careful categorisation and classification, to ensure proper understanding of the strengths and limitations of the specific tools both by teachers and learners. The selection of a particular resource for an LSP context, according to Aston (1996), should take into account and balance the following three criteria: "the extent to which the texts contained in a corpus should be limited to the particular domain of LSP (corpus specificity); the number and length of the text samples it should contain (corpus size); and the extent to which these mirror the universe of discourse which the corpus aims to capture (corpus representativeness)."

Thus, the major dichotomies in corpus classification, according to researchers, are the following:

- representative/reference corpora (the Brown Corpus, the Lancaster-Oslo-Bergen Corpus, the British National Corpus), both extensive and balanced in terms of content, genre and text length; and monitor corpora (e.g., the Collins COBUILD Bank of English), which adopt the sheer size as the basis for the corpus authority (Tribble 1997; Gabrielatos 2005);
- large corpora (e.g., of 100,000,000 word size) typically aiming at a broad coverage of language categories, and small corpora (e.g., 20,000 words), far more specialised by topic, genre or both (Aston 1997);
- general corpora, which reflect a certain language in all its contexts of use, and specialised corpora, focusing only on a particular context or user,
- written or spoken language collections;
- general English or a geographical variety of the language (British, Indian, Singaporean, etc.);
- synchronic (recording language in its particular stage of development) or diachronic corpora (enabling historical analysis of language over time);
- monolingual (with samples of only one language), multilingual/comparable (containing the same text-types in different languages), or parallel (with the same texts translated into different languages);
- native speaker, non-native speaker and learner corpora, the latter being represented by the Louvain International Corpus of Learners' English, concerned with the problem of learner writing and demonstrating interlanguage problems (see Pravec 2002, for a comprehensive discussion of learner corpora available).

To this list, one could add another crucial distinction, the one into ready-made corpora resources, available for local or global use, unrestricted or limited to registered users, and teacher-made collections, compiled based on freely available Internet materials from a specific ESP/EAP domain and searched with the use of text analysis tools (e.g., Web Concordancer, http://www.edict.com.hk/concordance/ConcUpload.htm; ConcApp, http://www.edict.com.hk/PUB/concapp/; TextSTAT, http://www.niederlandistik.fu-berlin.de/textstat/; Simple Concordance Program, http://www.textworld.com/scp/; AntConc, http://www.antlab.sci.waseda.ac.jp/software.html). It is especially this final dimension, which, with the view of promoting teacher autonomy and assisting materials development to a greater extent, will be devoted much greater attention in the present paper.

4.1. Ready-made corpora resources

As Tribble (1997) notes, students need many corpora instead of one, to become effective language users in many different contexts. Gavioli and Aston (2001) support this view, claiming that one of the prerequisites for effective hands-on concordancing by learners is the

availability of different corpora, both spoken and written, specialized language, particular geographical and social varieties, as well as large mixed corpora such as the British National Corpus. This is essential to enable learners to "compare data from different sources, and to discuss language use in relation to different types of text, topic, and genre" (Gavioli & Aston 2001: 245).

The sample resources listed below may serve as a good starting point for language teachers to introduce the elements of corpus linguistics in their teaching. The list is by no means complete, and has been compiled to indicate possible types of corpora rather than actual examples (for more, also within other foreign languages, see http://www.sfb441.uni-tuebingen.de/c1/corpora.html or http://devoted.to/corpora):

1. English:
   - full versions of specialist corpora with unlimited access (MICASE Michigan Corpus of Academic Spoken English, http://www.hti.umich.edu/m/micase/; International Corpus of Learner English – Polish section, http://ifa.amu.edu.pl/~kprzemek/concord2advr/search_adv_new.html);
   - official demonstration versions of renowned corpora, usually with only basic keyword search facilities (British National Corpus, http://sara.natcorp.ox.ac.uk/lookup.html; Collins COBUILD Bank of English, http://www.collins.co.uk/Corpus/CorpusSearch.aspx);
   - full access custom-made interfaces to renowned corpora developed by researchers (British National Corpus, http://www.lextutor.ca/concordancers/concord_e.html and http://corpus.byu.edu/bnc/; Brown Corpus, http://www.lextutor.ca/concordancers/concord_e.html, http://www.edict.com.hk/concordance and http://www.ldc.upenn.edu/cgi-bin/ldc/textcorpus?doc=yes&corpus=BROWN; Lancaster-Oslo-Bergen Corpus, http://www.edict.com.hk/concordance);
   - corpora compiled of selected works of the English literature, such as *Alice in Wonderland, The Lord of the Rings*, *Call of the Wild* or Sherlock Holmes stories (Online Concordancer, http://www.lextutor.ca/concordancers/concord_e.html, Web Concordancer, http://www.edict.com.hk/concordance);
   - corpora composed of newspaper articles and television news transcripts, also based on current issues of online newspapers (Online Concordancer, http://www.lextutor.ca/concordancers/concord_e.html; Web Concordancer, http://www.edict.com.hk/concordance; Reuters Corpora, http://trec.nist.gov/data/reuters/reuters.html);
   - learner corpora (PICLE Polish International Corpus of Learner English, http://ifa.amu.edu.pl/~kprzemek/concord2advr/search_adv_new.html; Online Concordancer, http://www.lextutor.ca/concordancers/concord_e.html; Web Concordancer, http://www.edict.com.hk/concordance);
   - thematic corpora, e.g., of telephone conversations (http://www.ldc.upenn.edu/cgi-bin/lol/swb/speechcorpus?&corpus=swb), business letters (http://ysomeya.hp.infoseek.co.jp/), EU legislation (http://logos.uio.no/opus/), culinary, ecotourism, computer and environmental protection texts (http://www.nilc.icmc.usp.br/cortec/ibusca.php), European Parliament session transcripts (Europarl, http://people.csail.mit.edu/koehn/publications/europarl/).

2. French:
   - literature collections to be searched with external concordancers (Galica, http://gallica.bnf.fr/; ATHENA, http://un2sg4.unige.ch/athena/html/fran_fr.html; The

French Collection at the University of Victoria, http://etext.lib.virginia.edu/collections/languages/french/; Clicnet, http://clicnet.swarthmore.edu/litterature/sujets/XIX.html; Bibliotheque Nationale du Canada, http://collection.nlc-bnc.ca/e-coll-e/index-f.htm);

- corpora compiled of various texts with dedicated concordancing solutions (ABU: la Bibliothèque Universelle, http://abu.cnam.fr/; Concordancier Web, http://www.edict.com.hk/concordance/WWWConcappF.htm);
- corpora composed of selected literature texts (Concordancier français en-ligne, http://www.lextutor.ca/concordancers/concord_f.html);
- representative corpora (FRANTEXT, http://www.lib.uchicago.edu/efts/ARTFL/databases/TLF/; Concordancier Web, http://www.edict.com.hk/concordance/WWWConcappF.htm);
- thematic and specialist corpora, e.g., of speech (ELICOP/ELILAP/LANCOM, http://bach.arts.kuleuven.be/pmertens/test/pmquery.html and http://bach.arts.kuleuven.be/pmertens/corpus/search/s.html) or European Parliament transcripts (Europarl, http://people.csail.mit.edu/koehn/publications/europarl/);
- newspaper article collections (LEXIQUM, http://retour.iro.umontreal.ca/cgi-bin/lexiqum; Concordancier français en-ligne, http://www.lextutor.ca/concordancers/concord_f.html; Reuters Corpora, http://trec.nist.gov/data/reuters/reuters.html).

3. German:

- representative corpora (Kookkurrenzdatenbank CCDB, http://corpora.ids-mannheim.de/ccdb/; COSMAS, http://www.ids-mannheim.de/cosmas2/; DWDS, http://www.dwds.de/?corpus=1&opt=corpora);
- corpora of randomly selected Internet materials (Leipzig Corpora Collection, http://corpora.informatik.uni-leipzig.de/download.html);
- newspaper article collections (NEGRA Corpus, http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/negra-corpus.html; DWDS, http://www.dwds.de/?corpus=1&opt=corpora; Reuters Corpora, http://trec.nist.gov/data/reuters/reuters.html);
- thematic and specialist corpora, e.g., the language of the former DDR or the spoken language (DWDS, http://www.dwds.de/?corpus=1&opt=corpora); politics and tourism (http://www.tu-chemnitz.de/phil/InternetGrammar/), European Parliament transcripts (Europarl, http://people.csail.mit.edu/koehn/publications/europarl/), EU legislation (OPUS, http://logos.uio.no/opus/).

4.2. Teacher-made corpora using the Web as a corpus

One of the problems that teachers might have with using corpora is limited access. With the free online tools described above, the opportunities of didactic use have become greater; however, there is still the temptation to use the Web as a corpus in the search for an even greater number of instances of use. Rundell (2000) claims that the Internet "is not a corpus at all according to any of the standard definitions: what it is a huge ragbag of digital text, whose content and balance are largely unknown". The argument here goes that the Web is in no way balanced, with some text-types very well-represented, while others hardly present at all – thus, contrary to the corpora described above, it must not be treated as a representative sample of any language. Similarly, Fletcher (2001), himself an advocate of the judicious use of the Web as a corpus, points out to the following problems of "Webidence as Evidence":

"ephemeral content of dubious reliability; journalistic, commercial and personal texts of unknown authorship and authority abound; assertions are intermingled with and represented as established fact, and details of sources and research methodology are documented haphazardly at best." As Fletcher's argument continues, Internet domains are not the most trustworthy indication of provenance, many webpages contain mainly titles and captions rather than full text, finally, "sloppy spelling and careless language" appear to be a norm, especially in discussion groups.

On the other hand, Rundell (2000) notices the obvious advantages of using the Web as a corpus: much greater size, up-to-dateness and greater likelihood of containing relatively rare or novel lexical items (such as Rundell's 'Hitchcockian'), where the standard corpus search proves to be unsatisfactory. Also Fletcher (2001) notices the potential of the Web as a constantly expanding, self-renewing machine-readable body of linguistic data, much richer in current language usage, infrequent expressions, text genres and domains than even the biggest standard reference corpus. Elsewhere (Fletcher 2007), some more essential reasons for using the Web as a corpus are enumerated: freshness and spontaneity, completeness and scope, linguistic diversity, representativeness and free availability. This last factor can stimulate corpus linguists to the compilation of very large corpora, such as almost 2 billion words (Baroni and Kilgariff, 2006) or parallel corpora confronting as many as 11 languages (de Schryver, 2002).

Thus, removing some of the problems outlined above argues for the development of concordancing solutions that will be geared towards providing the user with all or most of the functions of concordancing software, but still with the possibility of using the Web in general or specific websites as a corpus. In this way, corpora resources become customised to particular teacher's needs, based on carefully selected websites of a given nature.

The technical solutions enabling teachers to use the Web as a corpus, extracting websites and adding them to corpora or browsing specific domains online, are multifold, ranging from the advanced search features of a widely accessible search engine to dedicated corpus creation solutions. The selection of a tool to use depends on a number of factors, such as, among others, familiarity of the user with corpus linguistics procedures and terminology, the amount of resources to be retrieved, or global (online) or local (off-line) use. Below there is an overview of the widely accessible tools of this kind, focusing only on the applications of immediate use by language teachers with little background in computational linguistics (for a more advanced discussion of the process, see Lüdeling *et* al. 2007).

Robb (2003) advocates the use of a widely-known search engine, Google, as a concordancer, with the Web as a corpus, to make up for the shortage of ready-made corpora, obviously bearing in mind the following shortcomings:

- limited searching possibilities, only to look for specific words or phrases, not word categories or inflected forms;
- lack of control over the educational level, nationality, or other characteristics of the creators of the utterances found;
- lack of register pre-selection;
- misleading frequency statistics provided by the search engine, taking into account common duplicate hits;
- the provision of only raw output, not in the usual KWiC (Key Word in Context) format, without the possibility to subsort on adjacent words or generate frequency/collocates lists.

Fletcher (2004) adds to that list lack of support for sophisticated pattern matching searches, dubious hitlist mechanism based on paid advertising or incoming links page ranking, varied

occurrence reports depending on the Internet traffic, inability to construct searches with the use of wildcards, accented characters or case-sensitivity.

In order to maximize the benefits gained from a search engine concordance query, Robb (2003) recommends careful selection of target sources, using trusted websites to authenticate language usage. One way to do that could be to use the advanced searching options to narrow down the possible pool of sites only to a specific domain (e.g., gov. or edu.), where one can reasonably expect educated usage, or use specific searching syntax to direct the search to the specific site/domain/user, excluding possible mismatches (see Robb, 2003, for a detailed description).

Another solution is Webcorp (http://www.webcorp.org.uk/), a set of tools using the Web as a corpus for concordance searches (Rundell 2000; Kehoe & Renouf 2002; Morley *et al.* 2003). In contrast to a standard search engine, the output is a proper concordance with the custom amount of surrounding context, with all concordance lines presented on a single results page with links to the original sites to get fuller context. WebCorp utilizes particular search engines to retrieve results, however, it actually visits each of the pages to extract concordance lines from them. Additional options of WebCorp, both in terms of query formulation (word filter, date filter, wildcards and pattern matching, domain specification) and display (customizable concordance span, optional display of full sentences, preferable number of concordance lines) make it more suitable for linguistic research than pure information retrieval. An interesting feature is post-processing, or working on the results to sort them or remove unwanted concordance.

WebCONC (http://www.niederlandistik.fu-berlin.de/cgi-bin/web-conc.cgi) is a effectively simple online form generating KWiC concordances based on Google exclusively (with the specification of site language and the number of sites to be retrieved), with no option of deciding which of the Google search hits will be used for a corpus. An alternative way of online corpus compilation is, similarly to other tools, specification of particular website domains for retrieval. The query interface supports regular expressions, while the output display gives the concordances together with website URLs and full text display, with the context (both left and right) set between the values of 25, 50, 75 and 100 words.

WebAsCorpus.org with its Web Concordancer (http://webascorpus.org/searchwac.html) allows web search for particular words or phrases specifying the language of the pages. The options include setting the amount of left and right context, the number of webpages to be processed, selecting countries and using predefined SafeSearch content filtering mechanism blocking webpages with questionable content. The output is divided into particular webpages, with the keyword highlighted in green, and basic information about the number of words and paragraphs in a webpage. More advanced search capabilities, including wildcards, alternatives and exclusion, are available in WebAsCorpus.org English Web Corpus from the same search interface (http://webascorpus.org/searchwc.html).

An interesting tool for using selected online resources as a corpus for linguistic enquiry is GlossaNet, http://glossa.fltr.ucl.ac.be/scripts/gtoday/gtoday.pl (see Fletcher 2004), termed as "Corpus Linguistics & information retrieval" tool. Contrary to previous tools allowing the specification of an individually selected website, GlossaNet is a web spider search engine that allows automatic retrieval of online editions of more than 100 newspapers in 12 languages. Selected newspapers can be used as sources of data to locate attestations of words or syntactic structures, subject to basic concordancing procedures (queries using Part-of-Speech tags and regular expressions, expressed as words or finite state graphs). The output is displayed as a standard KWiC concordance, with sorting options available, however, the retrieval of a full sentence context in many cases fails due to the fact that the original article is no longer available online/at the same URL.

Sketch Engine (http://www.sketchengine.co.uk/) is a web-based service (available for free for a 30-day trial) which takes as its input a corpus of any language with an appropriate level of linguistic mark-up. The text analysis tools comprise a concordancer, a word sketch program, grammatical relations and a distributional thesaurus. The output generated as a word sketch is a one-page automatic corpus-based summary of a word's grammatical and collocational behaviour (Kilgariff *et al.* 2004). The built-in WebBootCaT module (Baroni and Bernardini, 2004; Baroni *et al.*, 2006), allows users to build their own instant corpus, using the selected websites as sources of data, extracting keywords and handling terminology in any language. The corpus compilation procedure involves specifying seed words/loading a text file with seed words, selecting the language for the corpus, as well as personal settings for selecting URLs manually and POS-tagging the corpus automatically. The resultant corpus can be browsed online using the SketchEngine concordancer or downloaded locally as a .txt.file.

KWiC Finder (http://www.kwicfinder.com/KWiCFinder.html) is a search-engine-based research tool, mining the Web for the occurrences of particular words and displaying concordances. The tool enables users to create sophisticated queries which are submitted to the AltaVista search engine, to later retrieve and produce a KWiC concordance of online documents. KWiCFinder supports wildcards (not only basic * and ?, but also "exactly one" and "zero or one"), "sic option" forcing an exact match of lower-case and plain characters, allows the operators "before" and "after" together with the specification of the number of words to separate them. Search options involve manipulation of the scope of the search, the size and format of the report, and whether to save documents matching the user's search criteria on the hard drive. Search reports, on the other hand, allow efficient evaluation of online documents, enabling the user to switch from one line per citation concordance to table or paragraph layout with keywords highlighted. According to Fletcher (2004), "user-enhanced search reports can be saved as stand-alone HTML pages for sharing with students or colleagues, who in turn can annotate, supplement, save and share them."

A corpus building tool itself, TEXTStat (http://www.niederlandistik.fu-berlin.de/textstat/software-en.html) enables using .html files directly from the Internet as components of a custom-made corpus. The user specifies the domain (e.g., www.bbc.co.uk) together with the number for the level of subpages to be retrieved, which are collected by an in-built web spider and added to a corpus. Similarly, the news reader feature enables automatic retrieval of discussion group postings, which can be a useful source of data for more informal and speech-like discourse samples.

4.3. Teacher-made corpora using selected texts

A viable alternative for ready-made corpora resources can be custom-made collections, compiled by ESP/EAP teachers with the use of pre-selected texts in response to the specific requirements of a particular teaching context (Fletcher 2004; Lee & Swales 2006)[1]. Such 'do-it-yourself corpora' will be an indispensable solution when students' needs cannot be satisfied by the above-listed ready-made resources, when representative corpora contain relatively little coverage of specialist areas or text types (Tribble 1997), since even 100-million British National Corpus is "ill-equipped to meet the needs of translators working with very specialised texts and confronted with specific terminology" (Zanettin 2001), or when the teacher aims at enhancing the classroom with the language of a particular domain,

[1] Aston (2002) shows an intermediate approach – creating a customised subcorpus within the British National Corpus compiled of texts containing particular query terms, claiming that this solution exploits the benefits of ad-hoc corpora avoiding its drawbacks. However, due to financial constraints limiting the availability of BNC, this particular strategy is not explored in greater detail here.

geographical area or register. Even though a standard reference corpus like BNC may not be adequate for the needs of a particular domain, Aston (2001) advocates confronting hypotheses formulated based on DIY corpora with existing corpora, to add credibility to the process.

Elsewhere, Aston (2002) points out to such important features of home-made corpora as control, certainty of the content, stimulation of user's creativity, increasing critical awareness through trial and error of corpus compilation, finally, triggering discussion and leading to self-improvement. Contrary to the approach of the Web as a corpus, which emphasises larger size achieved via automatic retrieval at the expense of relevance, corpora made of carefully selected texts (available locally or globally) involve greater teacher's control over content.

A DIY web corpus (for translation purposes) can be characterized as follows (Zanettin 2001):

- it is a collection of Internet documents, or more precisely of web pages in HTML
- it is created ad hoc as a response to a specific text to be translated
- it is an open corpus. More material can be added as the need arises
- it is disposable (Varantola 2000) or virtual (Ahmad et al.1994). It is not destined to be part of a more permanent corpus, and can be disposed of as soon as the translation is completed. Copyright permissions are not required
- like "parallel texts" it can be either bilingual comparable or target monolingual.

Out of the above criteria, while holding to some extent true for LSP teaching, especially the disposability of an ad-hoc corpus needs to be replaced with reusability in the teaching context, as the purpose of the LSP teacher should be creating a set of resources that will maintain consistency of lexical coverage of the specialist area. Similarly, ad-hoc corpora may be equally well compiled of texts available locally.

The starting point for the process of corpus compilation, as Kilgariff *et al.* (2005) demonstrate, is to formulate a detailed corpus-design document, agreeing at target size, target proportions for different text types, basing on generally accepted factors (Atkins *et al.*, 1992), but modified according to local needs. Góźdź-Roszkowski and Witczak-Plisiecka (2005) point out to the need to make the decisions about the following factors: size, theme, text type, medium (oral or written), authorship, language (native speaker, non-native or learner), publication date. When analysing the issue of sources selection for corpus compilation, Curado (2006) quotes Hunston (2002: 16), saying that "the selection of sources should reflect the communicative exchanges that take place in the target context of research and work". In other words, the LSP corpus should constitute a balanced view of materials, ranging from formal writing (e.g., technical reports and instruction manuals) to informal conversational messages (discussion forum postings). Whistle (1999: 75) reports after Polezzi (1993) that a custom-made FL corpus needs to satisfy three basic requirements: "it must be based on the learners' needs; its size should be determined by the nature of the course and the level of the learners; it must be flexible allowing addition and modification."

Thus, to give a practical example, when compiling his teacher-made corpus for business English students, Curado (2006) used such specific subject areas as accounting, management, marketing, MIS and statistics, with each area accounted for in a similar number of words, and such text sources as textbooks, journal articles, e-discussions and reports, varying in number within the area but adding up to a similar total.

An interesting perspective, with significant potential for LSP teaching, may be offered by compiling parallel corpora, containing the same texts translated into various languages. While such efforts for as many as 11 languages may be difficult (see de Schryver 2002), the compilation of bilingual L1-L2 parallel corpus from a particular domain should be quite feasible, given the versatility of the following data sources.

Obviously, the selection of sources, their balance resulting in the representativeness of a corpus (Biber *et al.* 1998) or, on the contrary, the overrepresentation of a certain genre, text type or register are the result of the needs of the class and the didactic purposes for the exploitation of a mini-corpus. Thus, the teacher may decide to create a corpus for the English for archaeology class that will be as representative as possible, trying to balance subareas, text genres and levels of formality, if the general language development of LSP students is the main goal. On the other hand, with a clear focus shared by most students in a group, e.g., Polish drivers planning to work in the public transportation system in the UK and Ireland, the teacher might use the sources more focused on the specific area of interest, such as the websites of drivers' trade unions, traffic regulations, bus operation manuals, to get a home-made corpus which is narrower in scope but more targeted at specific needs of students.

The possible sources of texts for a teacher-made corpus may be as follows:

- a CD-ROM (Microsoft Encarta, Hutchinson's, Grolier's) or online encyclopedia (Wikipedia, http://en.wikipedia.org), with texts browsable by topic and categorized into domains, for semi-formal register and essay/biography/process description genres (Tribble 1997);
- a legislation repository (European Commission's Eur-Lex, http://eur-lex.europa.eu/), with acts, treaties and agreements browsable by subject, date, keyword and the like, for highly formal register and act/contract/agreement genres;
- an archive of a subject matter discussion group, e.g., hosted at Yahoo!Groups (http://groups.yahoo.com), browsable by author, date and keyword, for informal register and written discussion genre;
- Frequently-Asked-Questions sections of ask-an-expert sites of the area (e.g., Refdesk.com's Ask the Experts, http://www.refdesk.com/expert.html; CIESE - Ask-An-Expert Links, http://www.k12science.org/askanexpert.html; Pitsco's Ask an Expert Site, http://www.askanexpert.com or Expert Central, http://expertcentral.com), for process descriptions and advice giving;
- a specific Web-based journal ('expert' writing – Lee & Swales 2006), retrievable via one of the many journal-finding interfaces (e.g., EBSCO – http://search.ebscohost.com/), characteristic for a particular area and well-renowned in the field, for formal register and article genre;
- M.A./Ph.D. dissertations in the area, retrieved from university websites or via EBSCO search interface, for formal register and research paper genre;
- official documents, manuals, procedure descriptions;
- a collection of institutional websites (e.g., airport authorities, see http://www.baa.com/, http://www.gtaa.com/, http://www.metwashairports.com/ - for more, see http://www.google.pl/search?q=airport+authority), for semi-formal register and company description genre.

Both Tribble (1997) and Curado (2006) draw the attention to the important step of technical editing raw texts before adding them to a corpus, by removing headers of discussion groups or converting the syntax of the opening sentence of encyclopedia entries to standard English, but not interfering with the language data. This was done only to prevent false conclusions following not from real language usage, but the requirements of a particular genre (or the "supergenre" of encyclopedia – Tribble 1997).

Another major issue to be addressed in the corpus compilation process is the awareness of copyright restrictions and the avoidance of potential copyright infringement. The point has been debated by some authors (Whistle 1999; Fletcher 2004), with the main line of reasoning being that the corpus compiler needs to obtain permission for educational use with no

modification or publication prospects, especially justified by the fact that concordancing software basically extracts the statements from their original context, which may be viewed by some as illegitimate use. However, Fletcher (2004) notes that "in the United States, a KWiC concordance of webpages appears to fall under the fair-use provisions of copyright law as well," while de Schryver (2002) adds that "as long as the compiled corpora are thus solely manipulated for research purposes, and are not used or published commercially, linguists should be on the right track." However, specific regulations may vary from country to country, and need to be checked beforehand to ensure the legality of the actions taken, and most probably the approach adopted by Kilgariff *et al.* (2005), namely contacting numerous potential text-donors, sending a short notice explaining the use of the text in the process, asking to contribute and sign permission letters, seems to be the safest possible.

5. Concordancing programs – a comparative review

A corpus alone, no matter how representative and authoritative it may be, is a collection of linguistic data of little use to both linguists and teachers without appropriately powerful tools enabling its querying for occurrences of words (concordances) in a variety of modes, simpler and more sophisticated, based on regular expressions or Part-of-Speech tags. A text analysis program like that gives the user the opportunity to formulate queries based on various schemes and strategies so that the output retrieved from a corpus allows verification of language hypotheses (Thomas 2002).

Apart from concordancing interfaces that existing corpora are equipped with, both officially available and custom-made by scholars, there is a plethora of open source or freeware concordancing programs to be used with any collection of texts gathered into a corpus. What needs to be noted here is that the teacher has greater flexibility and freedom in using the procedures of corpus linguistics with such programs, resulting in increased teacher autonomy and specialist language awareness.

Concordancing programs can be classified according to a variety of criteria, some of which are highlighted in Table 1 below. The initial step for in-class concordancing is the process of corpus compilation, where the programs will either demand a single file or will need various files to be added to a single corpus. The latter solution is more beneficial, as the user can make finer distinctions of meaning referring to the source file to analyse text register or genre, or can compile separate text subcorpora according to various criteria. On the technical side, different solutions demand various filetypes, which may call for converting files or copying text from a pdf file and pasting it to a txt file. What is to be noted is that in some tools (TextSTAT) the user can automatically retrieve selected Internet domains or discussion group posts into a corpus, thus speeding up the process significantly.

The most practical aspect of the process is searching, ranging from simple keyword search, through associated words search with multiple words input, searching a corpus for occurrences of words from a pre-defined list ('stoplist'), sorting concordancing by left or right collocates together with frequency information, finally, providing word and letter frequencies from the entire text. The searching process is intricately connected with the concordance output, where on the one hand the user can evoke the context of only the very sentence (Web Concordancer), retrieve full text with the keyword highlighted (TextSTAT), or gain access to the entire source file for closer inspection (antConc). The use of wildcards and regular expressions, part of word (prefix/suffix) search or case-sensitive search further enable the user to narrow down the output to the desired examples.

Most probably one of the crucial points in the consideration of concordancing solutions is their accessibility, expressed on the one hand in terms of price, but also conditioned by the

amount of linguistic or computational knowledge required to operate the program. Thus, one can distinguish commercial solutions (such as WordSmith Tools, http://www.lexically.net/wordsmith/ or Sketch Engine, http://www.sketchengine.co.uk); freely available online concordancing forms (Web Concordancer, http://www.edict.com.hk/concordance/) with menu items used to choose searching options; and freeware/shareware/open source text analysis software, downloadable from the Web for installation on an unlimited number of workstations (AntConc, ConcApp, TextSTAT, Simple Concordance Program).[2] Finally, one can reflect on the use of familiar office applications (word-processors or spreadsheets) for basic concordancing procedures (see Boxall 2006).

The other aspect of accessibility is expressed in such technological considerations as trouble-free operation, transparent interface, easy installation, intuitive icons, PC/Mac versions and compatibility with various operating systems. Finally, the metalanguage and linguistic knowledge necessary to formulate queries influence the accessibility of the tool, and one can distinguish simple concordancing solutions, applicable also for students with proper in-class training (Web Concordancer, ConcApp or TextSTAT) and more sophisticated software demanding greater linguistic knowledge in formulating precise queries (AntConc).

The factors summarised above have been used to subject five selected concordancing solutions to a comparative review. The choice of criteria (inspired by Ioannou-Georgiou 2002; Diniz 2005 and Davies 2005) was purposefully conditioned by the factors of interest to an average language teacher, without going into more sophisticated problems of corpus linguistics. All the programs are freely available solutions, with no limitations on installation or output display. They are similar in that all demand a teacher-made corpus to be searched, however, differ significantly within other areas. The order of the presentation in the table below, left to right, reflects the growing sophistication of the programs. However, it was not the purpose of the review to make a definitive judgment on one of the tools characterised, but rather juxtapose the most crucial factors outlined above for users to be able to select different solutions based on varied needs.

Web Concordancer, http://www.edict.com.hk/concordance/ConcUpload.htm
ConcApp, http://www.edict.com.hk/PUB/concapp/
TextSTAT, http://www.niederlandistik.fu-berlin.de/textstat/
Simple Concordance Program, http://www.textworld.com/scp/
AntConc, http://www.antlab.sci.waseda.ac.jp/software.html

Table 1. A comparative review of selected concordancing programs.

	Web Concordancer	**ConcApp**	**TextSTAT**	**Simple Concordance Program**	**AntConc**
AVAILABILITY					
Price	Free	Free	Free	Free	Free
Downloaded from the Web	No (online form)	Yes	Yes	Yes	Yes
Licence	Not applicable	Freeware	Freeware	Freeware	Freeware
Upgrade	Not applicable	Yes	Yes	Yes	Yes
Operating systems	Not applicable	Windows 3.1, 95,	Windows (no version	Windows 95+ Mac OS 10	Windows 98+,

[2] Widespread availability and unlimited access were the threshold conditions for the selection of concordancing tools for a comparative review. This is the only reason why some well-established concordancing solutions such as WordSmith Tools or MicroConcord did not get their due attention in this paper.

		98+	listed) Mac OS 10 Linux (when equipped with Python)		Mac OS 10.4.5., Vine Linux
PC/Mac versions	Not applicable	PC only	PC and Mac	PC and Mac	PC and Mac
Internet access	Yes	No	Yes (to retrieve websites for a corpus)	No	No
Interface language	EN	EN	EN, DE, DU, PT	EN	EN
TECHNICAL CRITERIA					
Hardware requirements	Not applicable	Standard	Standard	Standard	Standard
Installation	Not applicable	Easy	Easy	Easy	No installation
Reliability of operation	Yes	Yes	Yes	Yes	Yes
Navigation	Clear	Clear	Clear	Clear	Medium
User support systems	Sample tasks, procedure description	Standard Windows help, online tutorial	Link to a website with examples of regular expressions	Standard Windows help	No help
Interface transparency	Yes	Yes	Yes	Medium	Medium
Icons	No icons (drop-down menus)	Icons easy to understand	Non-standard icons	Transparent but few icons	No icons, menu items instead
Encoding	ASCII	ANSI, Unicode	All systems	No information	All systems
CORPUS COMPILATION					
Filetypes supported	txt	txt, rtf	doc, sxw, txt, html	ans, asc, txt	txt, html, htm, xml
Corpus compiled into one file	Yes	No	No	No	No
Retrieving Internet websites to a corpus	No	No	Yes	No	No
Retrieving discussion group posts to a corpus	No	No	Yes	No	No
Removing selected files from a corpus	No	Yes	Yes	Yes	Yes
SEARCHING PROCEDURE					
Keyword search	Yes	Yes	Yes	Yes	Yes
Associated words search	Yes	Yes	Yes	No	Yes
Available number of associated words	1	3	1	0	Unlimited
Wildcards use in search	No	No	Yes	No	Yes
Case-sensitive search	No	No	Yes	No	Yes
Regular expressions search	No	No	Yes	No	Yes
Part of word search (prefix, suffix search)	Yes	Yes	No	Yes	Only prefix
Specifying associated word distance	Yes (for sorting only)	Yes	Yes	No	Yes
CONCORDANCE OUTPUT					
KWiC (Key Word in Context) display	Yes	Yes	Yes	Yes	Yes
Keyword gapped display	Yes	Yes	No	No	No
Exporting concordance list as a file	No	No	Yes (txt, doc)	No	Yes (txt)
Exporting frequency list as a file	No	No	Yes (csv, xls)	No	No

Displaying results with reference to the source file	Not applicable – due to single file support	Yes	Yes	No	Yes
Displaying full context (full source text file)	No	Yes	Yes	No	Yes
Sorting by left/right collocates	Yes	Yes	Yes	No	Yes
Special characters table	No	No	No	Yes	No
ADDITIONAL FEATURES					
Frequency list	Yes	Yes	Yes	Yes	Yes
Word profile	No	Yes	No	Yes	No
Letter frequency	No	No	No	Yes	No
Concordance-based interactive test	No	Yes	No	No	No
Concordancing a text based on a word list (stoplist)	No	Yes	No	No	Yes

6. Conclusion

Together with the implementation of Information and Communication Technology in language teacher training curricula, there arises a need to address the issues expanding upon the teacher's inventory, empowering teachers to gain greater awareness of the target language as the object of study. An example of such an area is corpus linguistics procedures, which, when implemented into teacher training, should find their reflection later in teachers' more informed choices about the teaching content and resulting learner autonomy stimulated by discovery learning procedures.

The prerequisite for teacher-made concordancing, be it in general EFL/ESL or specialised areas, is the availability of corpora providing adequate coverage of genres, modes and topics. In many teaching contexts the wide range of ready-made corpora, as summarised in the present paper, will prove adequate for a number of language needs. However, due to an ever-changing nature of language and more and more crystallised language expectations of students, there may arise a necessity to create custom-made collections based on either online materials (Web as a Corpus), teacher-selected or learner-produced texts. Such an approach enables achieving greater relevance of the materials especially for the LSP contexts, as the ready-made corpora available for searching do not prominently represent specialist areas. With the proliferation of tools of varied complexity as demonstrated in the present paper, the implementation of teacher-made concordancing based on custom corpora should be available in most teaching contexts.

References

Aston, G. 1996. "What corpora for ESP?" *A Conference on ESP at the University of Pavia*, November 1996. [http://www.sslmit.unibo.it/~guy/pavesi.htm]

Aston, G. 1997. "Small and large corpora in language learning". J. Melia and B. Lewandowska-Tomaszczyk, eds. *PALC '97 Proceedings. The First International Conference: Practical Applications in Language Corpora (1997), University of Łódź, Poland*. Łódź: Łódź University Press. 51-62. [http://www.sslmit.unibo.it/~guy/wudj1.htm]

Aston, G. 2001. "Text categories and corpus users: a response to David Lee". *Language Learning & Technology* 5 (3): 73-76. [http://llt.msu.edu/vol5num3/aston/default.html]

Aston, G. 2002. "The learner as corpus designer". B. Kettemann and G. 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. 9-25.

Atkins, S., J. Clear, and N. Ostler. 1992. "Corpus design criteria". *Literary and Linguistic Computing 7* (1): 1-16.

Barlow, M. 2002. "Corpora, concordancing, and language teaching". *2002 KAMALL International Conference*. Daejon, Korea.

Baroni, M., and S. Bernardini. 2004. "BootCaT: Bootstrapping corpora and terms from the Web". *Proceedings of LREC 2004*: 1313-1316. [http://sslmit.unibo.it/~baroni/publications/lrec2004/bootcat_lrec_2004.pdf]

Baroni, M., and A. Kilgarriff. 2006. "Large linguistically-processed web corpora for multiple languages." *Proceedings of EACL, Trento, Italy*. [http://acl.eldoc.ub.rug.nl/mirror/E/E06/E06-2001.pdf]

Baroni, M., A. Kilgariff, J. Pomikálek, and P. Rychlý. 2006. "WebBootCaT: a web tool for instant corpora". *Proceedings of the EuraLex Conference 2006*. 123-132. [http://sketchengine.co.uk/trac/attachment/wiki/WBC/DocsIndex/webbootcat_eamt06.pdf?format=raw]

Biber, D., S. Conrad, and R. Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Boxall, D.R. 2006. "Do-it-yourself concordancing". A message sent to TESLCA-L discussion list, March 29, 2006.

Chen, Y.H. 2004. "The use of corpora in the vocabulary classroom". *The Internet TESL Journal* 10 (9). [http://iteslj.org/Techniques/Chen-Corpora.html]

Cobb, T. 1998. "Breadth and depth of lexical acquisition with hands-on concordancing". *Computer Assisted Language Learning* 12 (4): 345-360.

Cobb, T. 1999. "Applying constructivism: A test for the learner as scientist". *Educational Technology Research & Development* 47 (3): 15-31.

Crystal, D. 1991. *A Dictionary of Linguistics and Phonetics*. Oxford: Blackwell.

Curado Fuentes, A. 2006. "A corpus-based focus on ESP teaching". *Teaching English with Technology* 6 (4). [http://www.iatefl.org.pl/call/j_esp26.htm]

Davies, G. 2005. "ICT4LT Project: Evaluation forms". [http://www.ict4lt.org/en/evalform.doc]

Diniz, L. 2005. "Comparative Review: *TextSTAT 2.5, Antconc 3.0,* and *Compleat Lexical Tutor 4.0*". *Language Learning & Technology* 9 (3): 22-27. [http://llt.msu.edu/vol9num3/review2/default.html]

Fletcher, W.H. 2001. "Concordancing the Web with KWiCFinder". *American Association for Applied Corpus Linguistics Third North American Symposium on Corpus Linguistics and Language Teaching*, Boston, MA, 23-25 March 2001. [http://www.kwicfinder.com/FletcherCLLT2001.pdf]

Fletcher, W.H. 2004. "Facilitating the compilation and dissemination of ad-hoc Web corpora". G. Aston, S. Bernardini and D. Stewart, eds. *Papers from the Fifth International Conference on Teaching and Language Corpora*. Amsterdam: Benjamins. [http://kwicfinder.com/Facilitating_Compilation_and_Dissemination_of_Ad-Hoc_Web_Corpora.pdf]

Fletcher, W.H. 2007. "Concordancing the Web: Promise and problems, tools and techniques". M. Hundt, N. Nesselhauf, and C. Biewer, eds. *Corpus Linguistics and the Web*. Amsterdam: Rodopi. [http://kwicfinder.com/FletcherConcordancingWeb2005.pdf]

Gabrielatos, C. 2005. "Corpora and language teaching: Just a fling or wedding bells?" *TESL-EJ* 8 (4). [http://writing.berkeley.edu/TESL-EJ/ej32/a1.html]

Gavioli, L., and G. Aston. 2001. "Enriching reality: language corpora in language pedagogy". *ELT Journal* 55 (3): 238-246.

Godwin-Jones, B. 2001. "Tools and trends in corpora use for teaching and learning". *Language Learning & Technology* 5 (3): 7-12. [http://llt.msu.edu/vol5num3/emerging/]

Góźdź-Roszkowski, S., and I. Witczak-Plisiecka. 2005. Korpusy a języki specjalistyczne [*Corpora and languages for specific purposes*]. B. Lewandowska-Tomaszczyk, ed. *Problemy językoznawstwa korpusowego*. Łódź: Łódź University Press. 174-200.

Hasselgård, H. 2001. "Corpora and their use in research and teaching". [http://folk.uio.no/hhasselg/UV-corpus.htm]

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

Ioannou-Georgiou, S. 2002. "Selecting software for language classes". *Modern Language Teacher* 11 (3): 63-68.

Johns, T. 1988. "Whence and whither classroom concordancing?" T. Bongaerts, P. de Haan, S. Lobbe, and H. Wekker, eds. *Computer Applications in Language Learning*. Dordrecht: Foris Publications. 9-33.

Kehoe, A., and A. Renouf. 2002. "WebCorp: applying the Web to linguistics and linguistics to the Web". *WWW2002 – The Eleventh International World Wide Web Conference*, Honolulu, May 7-11 2002. [http://www2002.org/CDROM/poster/67/index.html]

Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. Harlow: Addison Wesley Longman.

Kilgariff, A., and G. Grefenstette. 2003. "Web as Corpus. Introduction to the special issue on the web as corpus". *Computational Linguistics* 29: 333-347.

Kilgariff, A., M. Rundell, and E. Uí Dhonnchadha. 2005. "Corpus creation for lexicography". [http://www.kilgarriff.co.uk/Publications/2005-KilgRundellUiD-Asialex-NCI.doc]

Kilgariff, A., P. Rychlý, P. Smrz, and D. Tugwell. 2004. "The Sketch Engine". Williams, Vessier, eds. *The 11th International EURALEX Congress*. UBS Lorient, France. 105-116. [http://www.sketchengine.co.uk/trac/attachment/wiki/SkE/DocsIndex/sketch-engine-elx04.pdf?format=raw]

Krieger, D. 2003. "Corpus linguistics: What it is and how it can be applied to teaching". *The Internet TESL Journal* 9 (3). [http://iteslj.org/Articles/Krieger-Corpus.html]

Krishnamurthy, R. 2001. "Learning and teaching through context - A data-driven approach". *TESOL Spain Newsletter* 24, 2001. [http://www.developingteachers.com/articles_tchtraining/corpora1_ramesh.htm]

Lee, D., and J. Swales. 2006. "A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora". *English for Specific Purposes* 25: 56-75.

Leech, G. 1997. "Teaching and language corpora - a convergence". A. Wichmann, S. Fligelstone, T. McEnery, and G. Knowles, eds. *Teaching and Language Corpora*. New York: Longman. 1-23.

Lüdeling, A., S. Evert, and M. Baroni. 2007. "Using Web data for linguistic purposes". M. Hundt, N. Nesselhauf, and C. Biewer, eds. *Corpus Linguistics and the Web*. Amsterdam: Rodopi. 7-24.

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

Morley, B., A. Renouf, and A. Kehoe. 2003. "Linguistic research with XML/RDF-aware WebCorp tool". [http://www2003.org/cdrom/papers/poster/p005/p5-morley.html]

Partington, A. 1998. *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam: John Benjamins.

Polezzi, L. 1993. "Concordancing and the teaching of *ab initio* Italian Language for Specific Purposes". *ReCALL* 9: 14-19.

Pravec, N.A. 2002. "Survey of learner corpora". *ICAME Journal* 26: 81-114.

Robb, T.N. 2003. "Google as a quick 'n dirty corpus tool". *TESL-EJ* 7 (2). [http://www-writing.berkeley.edu/TESL-EJ/ej26/int.html]

Rundell, M. 2000. "The biggest corpus of all". *Humanising Language Teaching* 2 (3). [http://www.hltmag.co.uk/may00/idea.htm]

de Schryver, G.M. 2002. "Web for/as corpus: a perspective for the African languages". *Nordic Journal of African Studies* 11 (2): 266-282.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. 1996. "EAGLES: Preliminary recommendations on corpus typology. Expert Advisory Group on Language Engineering Standards Guidelines EAG-TCWG-CTYP/P". [http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html]

St. John, E. 2001. "A case for using a parallel corpus and concordancer for beginners of a foreign language". *Language Learning & Technology* 5 (3): 185-203. [http://llt.msu.edu/vol5num3/stjohn/default.html]

Stevens, V. 1995. "Concordancing with language learners: Why? When? What?" *CAELL Journal* 6 (2): 2-10. [http://www.eisu.bham.ac.uk/johnstf/stevens.htm]

Thomas, J. 2002. "A ten-step introduction to concordancing through the Collins Cobuild Corpus Concordance Sampler". [http://web.quick.cz/jaedth/Introduction%20to%20CCS.htm]

Thomas, J. 2003. "Extending vocabulary knowledge with computers". *Teaching English with Technology* 3 (2). [http://www.iatefl.org.pl/call/j_tech13.htm#aword2]

Tribble, C. and G. Jones. 1990. *Concordances in the Classroom: A Resource Book for Teachers*. London: Longman.

Tribble, C. 1997. "Improvising corpora for ELT: Quick-and-dirty ways of developing corpora for language teaching". J. Melia and B. Lewandowska-Tomaszczyk, eds. *The First International Conference: Practical Applications in Language Corpora (1997), University of Łódź, Poland. PALC '97 Proceedings*. Łódź: Łódź University Press. [http://www.ctribble.co.uk/text/Palc.htm]

Weber, J.-J. 2001. "A concordance- and genre-informed approach to ESP essay writing". *ELT Journal* 55 (1): 14-20.

Whistle, J. 1999. "Concordancing with students using an 'Off-theWeb' corpus". *ReCALL* 11 (2): 74-80.

Zanettin, F. 2001. "DIY corpora: the WWW and the translator". B. Maia, H. Haller, and M. Urlrych, eds. *Training the Language Services Provider for the New Millennium.* Porto: Facultade de Lettras, Universidade do Porto. 239-248. [http://www.federicozanettin.net/DIYcorpora.htm]