МИНИСТЕРСТВО ВЫСШЕГО И СРЕДНЕГО СПЕЦИАЛЬНОГО ОБРАЗОВАНИЯ РЕСПУБЛИКИ УЗБЕКИСТАН

БУХАРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Факультет иностранных языков

Кафедра английского языкознания

«Рекомендуется к защите»

Декан факультета:

_____М.М.Жураева

«_____»_____2019 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

на тему: Significance of language corpora in teaching English

студента(ки) 4 курса_ Тоштемирова Дилдора Абдураим кизи по направлении 5120100 - Филология и обучение языкам (английский язык)

Научный руководитель: Турсунов Мирзо Ма

Рецензент: ХоликоваН.Н.

Бухара – 2019

CONTENTS

INTRODUCTION ……..…2-5

CHAPTER I. A LANGUAGE CORPUS AS A TOOL TO TEACH A FOREIGN LANGUAGE…….….5-26

1.1 General characteristic features of corpora…….....**.....**5-8

1.2 Various types and usages of language corpora……8-16

1.3 Learner corpora and interlanguage analysis ……..16-21

1.4 Challenges of using corpora in language teaching process……21-26

CHAPTER II. PRACTICAL USE OF LANGUAGE CORPORA IN THE PROCESS OF TEACHING A FOREIGN LANGUAGE.....27-51

2.1 The use of language corpora in teaching English ……..27-39

2.2 Using corpora for teaching lexis and collocation ……..39-43

2.3 Corpora as a tool of student autonomy and independence…….43-50

CONCLUSION……..51-53

A list of used literature……..54-55

Introduction

Today, in Uzbekistan a lot of reforms in every field are being done by our government which will be foundation stone of a strong, democratic state. For achieving these dreams each citizen of our country should add his or her contribution, strength, it is a holy duty of each of us.

Education system is also being reformed nowadays. New model in education will bring up young generation with strong ideas, people who have developed logical thinking to our society.

As our independent country has started to have diplomatic relationships with other independent states, a great demand of knowing foreign languages rose. Today Uzbekistan is going to begin teaching foreign languages with a new curriculum which can fit to our all demands, that is European standard of English knowing. According to new curriculum classes of English are funny, practical and interesting, full of different interactive methods which can let students be active, communicative, and creative during classes. Also, writing and reading language skills are considered to be very essential, as they often mark the level of a learner of a foreign language. Modern methodology demands all language skills to be integrated into one single unit or machine.

Actuality of the graduate qualification work: as today Uzbekistan is making a lot of reforms in education system, especially in learning foreign languages, it is actual to study every new sphere and method of new system.

Methodology has become an essential part of teaching and learning foreign languages, it is actual to teach these languages through modern, new, novice approaches, methods, techniques. Linguistics has presented methodology new aspects of language issues that have become an important part of teaching as well. Corpus linguistics as a methodology of linguistic research has gained such prominence over time that corpora have been used extensively in nearly all branches of linguistics. This work explores the potential uses of corpus data in

one of these areas – language teaching and learning. We will first discuss a wide range of issues related to using corpora in language pedagogy, including referencing publishing, syllabus design and materials development, language testing, teacher development, data-driven learner (DLL), teaching language for specific purposes, as well as learner corpus and interlanguage analysis. We are going to discuss the debate over the relevance of authenticity and frequency of corpora in language education as well as the future of corpus-based language pedagogy. It is actual to study new methodological basis of language teaching and learning related to linguistics.

The object of the graduate qualification paperis language corpora that is used in linguistics and methodology.

The subject of the graduate qualification paper is the process of foreign language teaching and learning.

The main aim of the graduate qualification paper is to find out the significance of language corpora in teaching English to pupils.

Tasks of the graduate qualification work:

- To find out information about language corpora;
- To relate language corpora with teaching foreign languages;
- To focus on different activities, methods and techniques used in foreign language classes;
- To focus on a number of factors which can be essential in teaching foreign languages based on language corpora;
- To focus on practical use of language corpora in teaching;
- To find and analyze some activities that can be useful to teach a foreign language implementing language corpora;

The level of studiedness of the graduate qualification work: foreign language skills have been studied by the following scientists and methodologists: Hunston, Osborne, Holones, Roger and others.

The novelty of the graduate qualification work:language corpora are quite new for our linguistics and methodology as well. Considering these we can state that not so many works are made based on language corpora in relation with methodology. We are going to study this issue from a new point of view.

Methodological basis and the way of investigation: There can be different methods to learn the theme: scanning, analyzing, commenting, comparing, concluding, interpreting, investigating, distinguishing, editing, translating and others.

The theoretical and practical value of the graduate qualification work: this graduation work can be useful nearly for every subject of a new curriculum as language corpora can be used in every class.That is to say, this final qualification work can be a good source of information both for teachers and students.

The contents of the graduate qualification work: The following graduate qualification work consists of an introduction, general characteristics of the work, two chapters; the first chapter has got four parts, the second contains three parts, a conclusion and a list of used literature, 55 pages.

CHAPTER I. A LANGUAGE CORPUS AS A TOOL TO TEACH A FOREIGN LANGUAGE

1.1 General characteristic features of corpora

Modern methodology has brought several new terms into language teaching and learning process. One of the essential issues in this sphere is the role of language corpora, its types and effective implementation in practice. Considering these features, we have decided to focus on the notion of corpora, its types and use in teaching English.

Corpora have been put to many different uses in fields as varied as natural language processing, critical discourse analysis and applied linguistics. Corpora are extensive digitized collections of written texts or transcribed oral material. This allows large quantities of authentic language data to be systematically analyzed. A variety of access routes for describing language, researching the acquisition of foreign and second languages, developing reference works, teaching and lesson-planning material, as well as for practical use during lessons, are opened up as a result.

A distinction is made between different types of corpus. There are text corpora consisting solely of written data, and so-called multi-modal corpora that also contain audio recordings and sometimes video as well. Usually additional data is available to complement the language data. This includes information about the author, place of publication or text type. It is possible to limit the corpus search, for instance to texts of a certain type or from a specified timeframe. The language data can be enriched with supplementary linguistic information (annotations). This includes information about the part of speech (POS tagging), about non-inflected basic forms (lemmatization) and the syntactical function of the individual words (parsing).

The use of corpora in language teaching and learning has been more indirect than direct. This is perhaps because direct use of corpora in language

pedagogy is restricted by a number of factors including, for example, the level and experience of learners, time constraints, curricular requirements, knowledge and skills required of teachers for corpus analysis and result interpretation, and the access to resources such as computers, and appropriate software tools and corpora, or a combination of these. Corpora are useful in several ways for lexicographers. The greatest advantage of using corpora in lexicography lies in their machine-readable nature, which allows dictionary makers to extract all authentic, typical examples of the usage of a lexical item from a large body of text in a few seconds. The second advantage of the corpus-based approach, which is not available when using citation slips, is the frequency information and quantification of collocation which a corpus can readily provide. Some dictionaries, e.g. COBUILD 1995 and Longman 1995, include such frequency information. Frequency data plays an even more important role in the so-called frequency dictionaries, which define core vocabulary to help learners of different modern languages, e.g. Davies for Spanish, Jones and Tschirner for German, Davies and de Oliveira Preto-Bay for Portuguese, Lonsdale and Bras for French, and Xiao, Rayson and McEnery for Chinese. Information of this sort is particularly useful for materials writers and language learners alike. A further benefit of using corpora is related to corpus markup and annotation. Many available corpora (e.g. the BNC) are encoded with textual (e.g. register, genre and domain) and sociolinguistic (e.g. user gender and age) metadata which allows lexicographers to give a more accurate description of the usage of a lexical item. Corpus annotations such as part-of-speech tagging and word sense disambiguation also enable a more sensible grouping of words which are polysemous and homographs. Furthermore, a monitor corpus allows lexicographers to track subtle change in the meaning and usage of a lexical item so as to keep their dictionaries up-to-date. Last but not least, corpus evidence can complement or refute the intuitions of individual lexicographers, which are not always reliable so that dictionary entries are more accurate. The above observations above are line with Hunston who summarizes the changes brought

about by corpora to dictionaries and other reference books in terms of five 'emphases': an emphasis on frequency, an emphasis on collocation and phraseology, an emphasis on variation, an emphasis on lexis in grammar and an emphasis on authenticity.

We should point out some types of corpora as well. The character of a corpus is determined by the type of texts that constitute it. Whereas Dr Johnson's corpus consisted largely of works by Shakespeare, Milton, Dryden and other literary figures, a modern general corpus will contain both written and transcribed spoken material from a wide range of media such as:

Books Emails Television Radio Conversations

Magazines Newspapers

Linguistic investigation will often require the analysis of specialized texts, and the corpora that a researcher uses or creates may reflect this. A corpus could, for example, consist entirely of any of the following:

- Samples of written US English
- Samples of spoken British English
- Business correspondence
- Legal contracts
- Old English
- Children's speech

Here are three other types of corpora that are worth mentioning: the learner corpus: Just as it sounds, this is a database of samples of English (or any language) that have been produced by learners. The writers of the Macmillan English Dictionary used such a corpus in order to identify the most common problems that learners experience when using English. The multilingual corpus: The three corpora that we have mentioned so far (Dr Johnson's, the World

English Corpus and the British National Corpus) are all monolingual in that they are made up entirely of texts in English. However, a corpus can consist of texts in two or more languages and provide translators with an effective tool for finding the equivalent ways in which different languages express similar ideas. Non-conventional corpora: If we define a corpus as nothing more than a large database of texts with a search facility, then we suddenly realize that most of us use corpora every day. If you have ever used the search window on the home page of onestopenglish to look for specific articles and/or lesson plans, then you have used the corpus principle. Similarly, the new Windows Live Hotmail has a search window which allows me to quickly locate previously received or sent emails as and when I need to (this has recently changed my life since my Inbox can be compared to an electronic version of Dr Johnson's garret). In fact, the biggest corpus of all is the World Wide Web itself and although it has not been specifically created for linguistic investigation, its usefulness for this purpose should by no means be disregarded.[1]Corpora can be a good assistant for researchers as well. One can make his work easier with the help of this teaching issue. The main purpose of corpora as a pedagogical tool is to focus on crucial language learning parts to form it in more comfortable shape to study.

1.2 Various types and usages of language corpora

A text corpus is a very large collection of text (often many billions of words) produced by real users of the language and used to analyse how words, phrases and language in general are used. It is used by linguists, lexicographers, social scientist, experts in natural language processing and in many other fields. A corpus is also be used for generating various language databases used in software development such as predictive keyboards, spell check, grammar

[1] Römer, U. 'Corpora and language teaching'. In A. Lüdeling and M. Kyto (eds.) Corpus Linguistics: An International Handbook, Berlin: Mouton de Gruyter, 2008, 129p.

correction, text/speech understanding systems, text-to-speech modules and many others.

Below we are going to focus on some types of text corpora. A text corpus can be classified into various categories by the source of the content, metadata, the presence of multimedia or its relation to other corpora. The same corpus can fall into more than one category if it fulfils the criteria for more categories.**Monolingual corpus.**Monolingual corpus is the most frequent type of corpus. It contains texts in one language only. The corpus is usually tagged for parts of speech and is used by a wide range of users for various tasks from highly practical ones, e.g. checking the correct usage of a word or looking up the most natural word combinations, to scientific use, e.g. identifying frequent patterns or new trends in language. Sketch Engine contains hundreds of monolingual corpora in dozens of languages.**Parallel corpus.** A parallel corpus consists of two monolingual corpora. One corpus is the translation of the other. For example, a novel and its translation or a translation memory of a CAT tool could be used to build a parallel corpus. Both languages need to be aligned, i.e. corresponding segments, usually sentences or paragraphs, need to be matched. The user can then search for all examples of a word or phrase in one language and the results will be displayed together with the corresponding sentences in the other language. The user can then observe how the search word or phrase is translated.**Multilingual corpus.** A multilingual corpus is very similar to a parallel corpus. The two terms are often used interchangeably. A multilingual corpus contains texts in several languages which are all translations of the same text and are aligned in the same way as parallel corpora. Sketch Engine allows the user to select more than two aligned corpora and the search will display the translation into all the languages simultaneously. When only two languages are selected, a multilingual corpus behaves as a parallel corpus. The user can also decide to work with one language to use it as a monolingual corpus.**Comparable corpus.** A comparable corpus is a set of two or more

monolingual corpora whose texts relate to the same topic.However, they are not translations of each other, and therefore, there are not aligned. When users search these corpora, they can use the fact, that the corpora also have the same metadata. An example of comparable corpora in Sketch Engine is CHILDES corpora or various corpora made from Wikipedia.**Learner corpus.** A learner corpus is a corpus of texts produced by learners of a language. The corpus is used to study the mistakes and problems learners have when learning a foreign language. Sketch Engine allows forlearner corpora to be annotated for the type of error and provides a special interface to search either for the error itself, for the error correction, for the error type or for a combination of the three options.**Diachronic corpus.** A diachronic corpus is a corpus containing texts from different periods and is used to study the development or change in language. Sketch Engine allows searching the corpus as a whole or only include selected time intervals into the search. In addition, there is a specialized diachronic feature called Trends, which identifies words whose usage changes the most of the selected period of time.**Specialized.** A specialized corpus contains texts limited to one or more subject areas, domains, topics etc. Such corpus is used to study how the specialized language is used. The user can create specialized subcorpora from the general corpora in Sketch Engine.**Multimedia.** A multimedia corpus contains texts, whichare enhanced with audio or visual materials or other type of multimedia content. For example, the spoken part of British National Corpus in Sketch Engine has links to the corresponding recordings which can be played from the Sketch Engine interface.

Corpora have revolutionized reference publishing (at least for English), be it a dictionary or reference grammar, in such a way that it is now nearly unheard of for new dictionaries and new editions of old dictionaries published from the 1990s onwards not to be based on corpus data, and ‘even people who have never heard of a corpus are using the product of corpus-based investigation’.[2]

[2] Kennedy, G., An Introduction to Corpus Linguistics. London: Longman, 1998, 95p.

Corpora are useful in several ways for lexicographers. The greatest advantage of using corpora in lexicography lies in their machine-readable nature, which allows dictionary makers to extract all authentic, typical examples of the usage of a lexical item from a large body of text in a few seconds. The second advantage of the corpus-based approach, which is not available when using citation slips, is the frequency information and quantification of collocation which a corpus can readily provide. Some dictionaries, e.g. COBUILD 1995 and Longman 1995, include such frequency information. Frequency data plays an even more important role in the so-called frequency dictionaries, which define core vocabulary to help learners of different modern languages, e.g. Davies for Spanish, Jones and Tschirner for German, Davies and de Oliveira Preto-Bay for Portuguese, Lonsdale and Bras for French, and Xiao, Rayson and McEnery for Chinese. Information of this sort is particularly useful for materials writers and language learners alike. A further benefit of using corpora is related to corpus markup and annotation. Many available corpora (e.g. the BNC) are encoded with textual (e.g. register, genre and domain) and sociolinguistic (e.g. user gender and age) metadata which allows lexicographers to give a more accurate description of the usage of a lexical item. Corpus annotations such as part-of-speech tagging and word sense disambiguation also enable a more sensible grouping of words which are polysemous and homographs. Furthermore, a monitor corpus allows lexicographers to track subtle change in the meaning and usage of a lexical item so as to keep their dictionaries up-to-date. Last but not least, corpus evidence can complement or refute the intuitions of individual lexicographers, which are not always reliable so that dictionary entries are more accurate. The above observations above are line with Hunston, who summarizes the changes brought about by corpora to dictionaries and other reference books in terms of five ‘emphases’: an emphasis on frequency, an emphasis on collocation and phraseology, an emphasis on variation, an emphasis on lexis in grammar and an emphasis on authenticity.

It has been noted that non-corpus-based grammars can contain biases while corpora can help to improve grammatical descriptions (McEnery and Xiao). The Longman Grammar of Spoken and Written English can be considered as a milestone in reference publishing. Based entirely on the 40-million-word Longman Spoken and Written English Corpus, the grammar gives 'a thorough description of English grammar, which is illustrated throughout with real corpus examples, and which gives equal attention to the ways speakers and writers actually use these linguistic resources'. The new corpus-based grammar is unique in many different ways, for example, by taking register variations into account and exploring the differences between written and spoken grammars.[3]

While lexical information forms, to some extent, an integral part of the grammatical description in Biber et al, it is the Collins COBUILD series, that focus on lexis in grammatical descriptions (the so-called 'pattern grammar', Hunston and Francis. In fact, Sinclair et al flatly reject the distinction between lexis and grammar. While pattern grammars focusing on the connection between pattern and meaning challenge the traditional distinction between lexis and grammar, they are undoubtedly useful in language learning as they provide 'a resource for vocabulary building in which the word is treated as part of a phrase rather than in isolation'.

In the dictionary family, perhaps the most important member as far as language pedagogy is concerned is a learner dictionary. Yet corpus-based learner dictionaries have a quite short history. It was only in 1987 that the Collins COBUILD English Dictionary was published as the first 'fully corpus-based' dictionary. Yet the impact of this corpus-based dictionary was such that most other publishers in the ELT market followed Collins' lead. By 1995, the new editions of major learner's dictionaries such as the Longman Dictionary of Contemporary English (LDOCE, 3rd edition), the Oxford Advanced Learner's

[3] Burnard, L. and McEnery, A. (eds.), Rethinking Language Pedagogy from a Corpus Perspective. New York: Peter Lang, 2000, 40p.

Dictionary (OALD, 5th edition), and a newcomer, the Cambridge International Dictionary of English (CIDE, 1st edition) all claimed to be based on corpus evidence in one way or another.

One of the important features of corpus-based learner dictionaries is that their inclusion of quantitative data extracted from a corpus. Another important feature, which is also related to frequency information, is that such dictionaries typically select the vocabulary used from a controlled set when defining the entry for a word. Producing definitions in an L2 that language learners can understand is a problem; language learners may not have a very well developed L2 vocabulary. This makes it necessary and desirable for dictionary makers to limit the vocabulary they use when defining words in a dictionary. Nowadays, most learner dictionary makers prepare a list of defining words, usually ranging from 2,000 to 2,500 words, based on the frequency information extracted from corpora as well as on the lexicographers' experience of defining words.

As noted earlier, an important use of corpus data for lexicography is in the area of example selection so that nowadays most dictionaries of English use corpora as the source of their examples. In the case of learner dictionaries, however, there was a tradition of using examples invented by lexicographers, rather than authentic materials, in dictionary production, because they believed that foreign language learners have difficulty understanding authentic materials and therefore have to be presented with simple, rewritten examples in which the use of a given word is highlighted to show its syntactic and semantic properties. It was corpus-based learner dictionary work which challenged this received wisdom. The COBUILD project broke with tradition and used authentic data extracted from corpora to produce illustrative examples for a learner dictionary. The use of authentic examples in learner dictionaries is an area where corpus-based learner dictionaries have innovated.

While indirect uses such as syllabus design and materials development are closely associated with what to teach, corpora have also provided valuable

insights into how to teach. Of Leech's three focuses, direct uses of corpora include 'teaching about', 'teaching to exploit', and 'exploit to teach', with the latter two relating to how to use. Given a number of restricting factors as noted in section 2, direct uses have so far confined largely to learning at more advanced levels, for example, in tertiary education, whereas in general English language teaching (let alone to mention other foreign languages), especially in secondary education, the direct use of corpora is 'still conspicuously absent'.

'Teaching about' means teaching corpus linguistics as an academic subject like other sub-disciplines of linguistics such as syntax and pragmatics. Corpus linguistics has now found its way into the curricula for linguistic and language related degree programs at both postgraduate and undergraduate levels. 'Teaching to exploit' means providing students with 'hands-on' know-how, as emphasized in McEnery, Xiao and Tono, so that they can exploit corpora for their own purposes. Once the student has acquired the necessary knowledge and techniques of corpus-based language study, learning activity may become student centered. 'Exploiting to teach' means using a corpus-based approach to teaching language and linguistics courses (e.g. sociolinguistics and discourse analysis), which would otherwise be taught using non-corpus-based methods. If the focuses of 'teaching about' and 'exploiting to teach' are viewed as being associated typically with students of linguistics and language programs, 'teaching to exploit' relates to students of all subjects which involve language study and learning, who are expected to benefit from the so-called data-driven learning (DDL) or 'discovery learning'.[4]

The issue of how to use corpora in the language classroom has been discussed extensively in the literature. With the corpus-based approach to language pedagogy, the traditional 'three P's' (Presentation – Practice – Production) approach to teaching may not be entirely suitable. Instead, the more

[4]Bernardini, S., Competence, Capacity, Corpora: A Study in Corpus-aided Language Learning. Bologna, 2000, 76p.

exploratory approach of 'three I's' (Illustration – Interaction – Induction) may be more appropriate, where 'illustration' means looking at real data, 'interaction' means discussing and sharing opinions and observations, and 'induction' means making one's own rule for a particular feature, which 'will be refined and honed as more and more data is encountered'. This progressive induction approach is what Murison-Bowie would call the interlanguage approach: namely, partial and incomplete generalizations are drawn from limited data as a stage on the way towards a fully satisfactory rule. While the 'three I's' approach was originally proposed by Carter and McCarthy to teach spoken grammar, it may also apply to language education as a whole, in our view.[5]

Teaching-oriented corpora are particularly useful in teaching languages for specific purposes (LSP corpora) and in research on L1 (developmental corpora) and L2 (learner corpora) language acquisition. Such corpora can be used directly or indirectly in language pedagogy as discussed in previous sections.In addition to teaching English as a second or foreign language in general, a great deal of attention has been paid to domain-specific language use and professional communication (e.g. English for specific purposes and English for academic purpose). For example, Thurstun and Candlin explore the use of concordancing in teaching writing and vocabulary in academic English. Hyland compares the features of the specific genres of meta-discourse in introductory course books and research articles on the basis of a corpus consisting of extracts from 21 university textbooks for different disciplines and a similar corpus of research articles. Upton and Connor undertake a 'moves analysis' in the business English using a business learner corpus. The authors approach the cultural aspect of professional communication by comparing the 'politeness strategies' used by learners from different cultural backgrounds. Thompson and Tribble examine citation practices in academic text. Koester argues, on the basis of an analysis of the performance of speech acts in workshop conversations, for

[5]Aijmer, K. Corpora and Language Teaching. Amsterdam: John Benjamins, 2009, 84p.

a discourse approach to teaching communicative functions in spoken English. Yang and Allison study the organizational structure in research articles in applied linguistics. Carter and McCarthy explore, on the basis of the CANCODE corpus, a range of social contexts in which creative uses of language are manifested. Hinkel compares the use of tense, aspect and the passive in L1 and L2 academic texts. Xiao reviews a number of case studies using domain specialized multilingual corpora to teach domain specific translation. Studies such as these demonstrate that LSP corpora are particularly useful in teaching language for specific purposes and professional communication.

1.3 Learner corpora and interlanguage analysis

Two kinds of corpora that emerged in the 1990s have not only greatly contributed to the vitality of corpus linguistics but have also revived contrastive analysis and interlanguage research. They are learner corpora and multilingual corpora.

The creation and use of learner corpora in language pedagogy and interlanguage research has been welcomed as one of the most exciting recent developments in corpus-based language studies. If native speaker corpora of the target language provide a top-down approach to using corpora in language pedagogy, learner corpora provide a bottom-up approach to language teaching.A learner corpus, as opposed to a "developmental corpus" composed of data produced by children acquiring their mother tongue (L1), comprises written or spoken data produced by language learners who are acquiring a second or foreign language. Data of this type has particularly been useful in language pedagogy and second language acquisition (SLA) research, as demonstrated by the fruitful learner corpus studies published over the past decade; and Myles 2005 for recent reviews). SLA research is primarily concerned with 'the mental

representations and developmental processes which shape and constrain second language (L2) productions'. Language acquisition occurs in the mind of the learner, which cannot be observed directly and must be studied from a psychological perspective. Nevertheless, if learner performance data is shaped and constrained by such a mental process, it at least provides indirect, observable, and empirical evidence for the language acquisition process. Note that using product as evidence for process may not be less reliable; sometimes this is the only practical way of finding about process. Stubbs draws a parallel between corpora in corpus linguistics and rocks in geology, 'which both assume a relation between process and product. By and large, the processes are invisible, and must be inferred from the products.' Like geologists who study rocks because they are interested in geological processes to which they do not have direct access, SLA researchers can analyze learner performance data to infer the inaccessible mental process of second language acquisition. Learner corpora can also be used as an empirical basis that tests hypotheses generated using the psycholinguistic approach, and to enable the findings previously made on the basis of limited data of a small number of informants to be generalized. Additionally, learner corpora have widened the scope of SLA research so that, for example, interlanguage research nowadays treats learner performance data in its own right rather than as decontextualized errors in traditional error analysis.[6]

At the pre-conference workshop on learner corpora affiliated to the International Symposium of Corpus Linguisticsheld at the University of Lancaster, the workshop organizers Yukio Tono and Fanny Meunier observed that learner corpora are no longer in their infancy but are going through their nominal teenage years – they are full of promise but not yet fully developed. In language pedagogy, the implications of learner corpora have been explored for curriculum design, materials development and teaching methodology (cf. Keck 2004: 99). The interface between L1 and L2 materials has been explored.

[6] Allan, Q. 'Enhancing the language awareness of Hong Kong teachers through corpus data'. Journal of Technology and Teacher Education 7/1, 1999, 20p.

Meunier (2002), for example, argues that frequency information obtained from native speaker corpora alone is not sufficient to inform curriculum and materials design. Rather, 'it is important to strike a balance between frequency, difficulty and pedagogical relevance. That is exactly where learner corpus research comes into play to help weigh the importance of each of these'. Meunier also advocates the use of learner data in the classroom, suggesting that exercises such as comparing learner and native speaker data and analyzing errors in learner language will help students to notice gaps between their interlanguage and the language they are learning. Interlanguage studies based on learner corpora which have been undertaken so far focus on what Granger calls 'Contrastive Interlanguage Analysis (CIA)', which compares learner data and native speaker data, or language produced learners from different L1 backgrounds. The first type of comparison typically aims to identify under or overuse of particular linguistic features in learner language while the second type aims to uncover L1 interference or transfer. In addition to CIA, learner corpora have also been used to investigate the order of acquisition of particular morphemes.[7] Readers can refer to Granger et al for recent work in the use of learner corpora, and read Granger for a more general discussion of the applications of learner corpora such as the International Corpus of Learner English (ICLE).In addition to SLA research, learner corpora can also be used directly in classroom teaching. For example, Seidlhofer and Mukherjee and Rohrbach demonstrate how a 'local learner corpus' containing students' own writings can be used directly for learning by coping with students' questions about their own or classmates' writings, or analyzing and correcting errors in such familiar writings.

We have so far discussed how corpora, including those teaching oriented corpora like LSP corpora and learner corpora, can be used directly or indirectly in language pedagogy. The section that follows seeks to demonstrate the predictive and diagnostic power of the integrated approach that combines

[7] Aston, G. (ed.) Learning with Corpora. Houston, TX: Athelstan, 2001, 163p.

contrastive corpus linguistics with interlanguage analysis in second language acquisition research as advocated in Römer, via a case study of passive constructions in Chinese learner English.[8]

Corpora analysis for language teaching and the role it plays in the teaching-learning processhas both advantages and disadvantages. This paper however is limited to showing some of theadvantages of using corpora to inform linguistic practices in the English classroom. It should benoted that the advantages mentioned below are just a small sample of a larger list.[9]

a) Corpora can inform deductive and inductive approaches to English teaching. In the deductive approach to teaching English, corpora analysis provides evidence thatinforms teachers (especially to those who are non-native speakers) about the use of languageelements they are presenting in class and provides them with clear and authentic examples of thelanguage elements.In the inductive approach to teaching English, corpora analysis provides students with datato infer language rules by themselves.

b) The use of corpora improves English programs and materialsdesigncourses. Gabrielatos suggests that teachers and material writers may unwittingly present theirpersonal informal observations about language as the true and full picture of language structureand use, or present their own preferred usage as the only 'correct' or 'acceptable' one'.However, because corpora informs us about the use of language elements in a way that ourintuition cannot, corpora can enlighten the syllabus, and most teaching materials can be basedaround corpus data. A clear example of this is the use of modals and the form in which they arepresented in English textbooks. According to the corpora-based studies on modals (e.g. would, can,might etc.) carried out by several researchers in which they compared how speakers and writers use this language to how

[8] Baker, M., 'Corpus linguistics and translation studies: implications and applications'. In M. Baker, G. Francis and E. Tognini-Bonelli (eds.) Text and Technology: in Honour of John Sinclair. Amsterdam: Benjamins, 1993, 45p.
[9] Ball, F., 'Using corpora in language testing'. Research Notes, 2001, 8p.

textbooksclaim students should use it. They found that textbooks are not teaching the full inventory of modallanguage, they are also providing confusing explanations for some of the language they teach. Hyland concludes that for the most part, modal expressions are simply introduced without system or comment and are summarily dealt with in a single exercise, which fails to emphasize either their function or importance.[10] Generally, the range of modal verbs addressed and the information provided on their use is inadequate… In the same vein, Williams affirms that the selection of examples is unclear, but she would suspect that authors too enthusiastically use introspection or a type of educated hunch, instead of an empirical research.Another example related to English teaching material design and its flaws is the honest review Swales wrote about his book Academic writing for Graduate Students, in which he recognizes the faults with the textbook syllabus. Swales and his co-author wrote the book before linguistic computerized analysis became a tool for linguists and material designers. They affirm that the book was written "before we became aware of corpora, and was based on our own lengthy experiences as writing instructors". Since Swales had no corpus data to inform his material, students are given a confusing picture of how imperatives (e.g. see, notice, suppose, consider) are used in academic writing. Swales et al. (1998) demonstrated this by comparing the imperatives in his textbook with the more common imperatives found in a corpus of academic writing that contained texts from different areas: art history, chemical engineering, communication studies, experimental geology, history, linguistics, literary criticism, philosophy, political science, and statistics. After the corpus analysis he concludes that from ten lexical choices; notice, imagine, refer, observe, take the case of, and disregard either did not occur in his corpus or occurred no more than twice in the main text. Swales et al. observe that verbs like suppose may occur in mathematical arguments, but with more frequency in non-mathematical philosophy. They do not mention see, and speculate that

[10]Mindt, D. 'English corpus linguistics and the foreign language teaching syllabus' in J. Thomas and M. Short (eds.) Using Corpora for Language Research, London: Longman, 1996, 94p.

consider is "probably rare outside (philosophical) arguments, whereas the current data suggest its common use in at least the major school of theoretical linguistics".[11]

1.4 Challenges of using corpora in language teaching process

It is often argued that corpora provide learners with 'authentic' or 'real' language, and since these words echo the key features of Communicative Language Teaching (CLT, hereafter) method that favors the use of authentic and real language over concocted ones, it is often assumed that corpus-based language materials are well-suited for CLT. However, some of the researchers have cast doubt on whether language data in corpora are truly authentic. Widdowson contrasted the concept of 'genuineness' and 'authenticity' and argued that 'genuineness' is the property of texts and is an absolute quality, while 'authenticity' is the characteristic of discourse interpretation. He claimed that language in corpora can be genuine, but it is not authentic because it is isolated from discoursal and communicative nature of language. In other words, language data in corpora do not tell us much about the authors of messages, their illocutionary intentions, the intended audience, and the circumstances in which the messages were produced. In line with Widdowson's argument, Cook emphasized the importance of context especially in spoken language and stated that "speech is often inseparable from [...] the circumstances of its production, and can only be apprehended in the context of the knowledge of the participants, their paralanguage and the situation". To cope with the problem of corpora in relation to their 'authenticity,' it was suggested that they should be pedagogically mediated and authentificated when they are intended to be used in the classroom. In other words, they must be re-contextualized in a pedagogical setting and be tuned to specific classroom purposes and thereby make them real

[11]Keck, C. 'Corpus linguistics and language teaching research: bridging the gap'. Language Teaching Research 8(1), 2004, 67p.

and relevant for learners. In this light, Gavioli and Aston stated that "the question is not whether corpora represent reality but rather, whether their use can create conditions that will enable learners to engage in real discourse, authenticating it on their terms". Seidlhofer also maintained that the meaning of authenticity that most corpus linguists pursue (i.e., language description) is different from its meaningin language pedagogy.[12]

A number of corpora in different languages have been created and their accessibility has been significantly improved, as they are readily available on the Web. In particular, a number of English corpora (e.g., BNC, COCA) have been created and their utility as an English teaching tool has been widely recognized. However, corpora seemed to have little impact on language pedagogy despite their potentials as a language teaching tool and resource. One of the fundamental problems stems from how corpora were created on what purposes. Indeed, mega-sized general corpora make a perfect sense to linguists who believe that the balance and representativeness of corpora are critical. For example, a large-sized corpus is essential in lexicography because they need to obtain a sufficient number of occurrences of lexical or structural items in order to compare a relative frequency of occurrences. However, they do not seem particularly tailored to language pedagogy. For example, learners often end up with hundreds of concordance lines that are sometimes messy, ambiguous, and even misleading, and then they might become quickly overwhelmed by too much data that is not directly relevant to their learning. In this vein, Gavioli warned that the use of corpora in the classroom can leave learners too much alone, overwhelmed by information and resources. A more serious problem resides in the content of corpora. The words and sentence structures that collocate with key words are often way beyond learners' level of linguistic competence, and the retrieved

[12] Mishan, F. Designing Authenticity into Language Learning Materials. Chicago: Chicago University Press, 2005, 54p.

instances are often from unfamiliar and widely differing contexts that make it more difficult for learners to interpret them.[13]

One of the distinguished merits of corpora is that one can access frequency data of lexical items using the built-in software and this can be substantially conducive to efficiency of learning. That is, by focusing on high frequency lexical items, learners can acquire the target language more efficiently, and this type of information is invaluable especially for EFL teachers and learners due to cost-effectiveness it can provide. In this light, some of the corpus linguists claimed that a high frequency of occurrence directly corresponds to a high effectiveness in language teaching and material development. However, this view has been criticized by other researchers. Widdowson contended that even though quantitative corpus findings are valuable not only for language description but also for language teaching, frequency data should not be automatically taken as the sole criterion for pedagogical decision. Language teachers should also consider other crucial factors such as relative easiness (learnability: for example, primacy of the concrete over the abstract) and how much core or nuclear (generative value) words and structures are, together with the learning context in which learners are situated (e.g., learners' proficiency level, teaching objectives, and curriculum, etc.).[14]

One of the major barriers that make learners or even teachers less accessible to corpora is that few teachers and learners are knowledgeable about the use of corpora. They are often neither familiar with built-in software nor experienced in dealing with corpus data that involve a range of linguistic and metalinguistic knowledge. Tribble argued that teachers do not seem to use corpora very much in their classrooms and it is mainly because most of them do not have extensive experiences with corpora. The problem appears to be associated with a few

[13] www.languagecorpora.com

[14] Biber, D., Conrad, S. and Reppen, R., Corpus Linguistics: Investigating Language Structure and Use. Cambridge: Cambridge University Press, 1998, 45p.

practical issues. First, even though the number of readily-available corpora on the Web has been increasing and their interface has been somewhat improved, their user-friendliness still fell short of the standard. For example, the existing part-of-speech tagging system is too much complex and specific for non-linguists (e.g., UCREL tag sets include about 150 different tags), and it is even inconsistent among the corpora (e.g., COCA vs. BNC),causing confusion. In addition, teachers themselves are not familiar with the process of the interpretation and categorization of corpus data by which the data can be analyzed. Indeed, corpora entail not only quantitative analyses of data (e.g., frequency of occurrences) but also qualitative ones in which human decisions should be involved (e.g., what to include and exclude and how to regulate and interpret data), which makes the analysis highly demanding and time-consuming. Furthermore, as some researchers noted, it is crucial for teachers to make corpus data 'palatable' for the learner by 'preselecting' and 'digesting' raw corpus data in order to make them pedagogically more relevant. In other words, the data should be often modified by selecting familiar contents, reducing the quantity of data, and simplifying the task, which can be another challenging task for teachers. Last but not the least, the problem has to do with the curricular requirements of school. All of the materials and activities used in the classrooms should be compatible with the curricular requirements and many of the teachers are pressed for time to cover the textbook following the school curricular. Thus, creating corpus-based materials and activities can be too much demanding and unmanageable especially when they are handling things that they are not familiar with.

How can one implement corpora into the language classroom successfully? The answer to the question appears not that simple and it may vary according to the context in which teachers and learners are situated. However, the successful integration of corpora into secondary school in Korea seems dependent on at least a few pedagogical conditions/requirements. Therefore, in

this section, taking into consideration some of challenges and limitations of existing corpora in language teaching, a few suggestions will be made for the integration of corpora into the classroom along with their rationales.

However, there are some probable solutions to this problem:

One of effective ways of using a local learner corpus is to make a comparison between learners' own language use and that of native speakers. This type of analysis can give information about qualitative (misuse) and quantitative differences (over- and underuse) in lexical, grammatical, and discourse features, helping teachers design and choose materials for their own students. As aforementioned, even though frequency is one of the most important criteria for pedagogical decision, it is necessary to strike a balance between frequency, difficulty and pedagogical relevance in the classroom. That is exactly where a local learner corpus comes into play to figure out the relative importance, identifying the forms which are problematic for particular learners. Research on native corpora would provide a native language description and be used when deciding the teaching agenda. However, research on a local learner corpus should be used to modify this agenda to meet the needs of the specific learner population. From learners' perspectives, while comparing their language use and that of native speakers, they can notice the gap between them and be prompted to realize what they do not know and what they should know, initiating a process of restructuring of their linguistic knowledge. Mukherjee contended that corpus-based DDL activities can raise learners' awareness of language by enabling them to analyze both negative evidence (provided by a learner corpus) and positive evidence (provided by a native corpus). For the reasons presented above, Korean EFL teachers in secondary school may need to develop a home-made, local learner corpus that consists of their own students' written/spoken texts. Then they can analyze the texts of their own students and reflect the result of analysis in material design and classroom activities. For example, DDL activities can be designed in such a way that learners are made

more aware of their common mistakes and that knowledge reconstruction may occur, leading to grammatical rehabilitation.

Indeed, the appropriate and effective use of corpora in the classroom is partly a technical issue, but primarily a pedagogical one. If the use of corpora in the classroom is not extensively discussed and researched to develop a pedagogical blueprint for the integration of corpora into the classroom, the purported educational outcome that a number of corpus linguists simply expected would not accrue to learners and teachers.

CHAPTER II. PRACTICAL USE OF LANGUAGE CORPORA IN THE PROCESS OF TEACHING A FOREIGN LANGUAGE

2.1 The use of language corpora in teaching English

We introduce the idea of using corpora – the linguist's name for 'big data' – in language research, and sketch its history, first in linguistics in general, then in language learning and teaching. We then take a careful look at the hazards of using corpora in language learning, and arrive at some maxims for when and how they have a place: firstly, don't scare the students; then, use the corpus when the dictionary does not tell you enough, and moreover, disguise the corpus as a dictionary. We then introduce Sketch Engine, and show how it implements these ideas through SKELL, its language-learner interface. We show how corpora can be used, both in the classroom, and in the background, for syllabus design, where we have used corpora of learner output to identify patterns of overuse and underuse, with implications for what needs teaching.

As we know big data is big news. It is the big new thing in science: newspapers run supplements on what it means and how it will change the world, and politicians announce initiatives and inaugurate new research centers so that their country's scientists will be leading the charge. So – how might that relate to language teaching? Within linguistics, the 'big data' movement antedates the name 'big data' and is called, rather, corpus linguistics. A 'corpus' is a 'body' of data, and linguists call their big datasets 'corpora' (the Latinate plural of 'corpus'). A corpus is a collection of pieces of language, so-called when used for language research, and it may be anything from newspaper articles, to transcripts of everyday conversations or chat shows or lectures, to novels or letters or advertising brochures or shopping lists.

Back in the 1980s, lexicographers and dictionary publishers were making the case for big data long before it was a familiar one. They knew, as James Murray, editor of the Oxford English Dictionary from 1879 to his death in 1915, had before them, that languages are big: for dictionaries to get good coverage of a language, and not to make embarrassing omissions, they need very, very large amounts of data, and tools to support finding all the words and phrases in them. James Murray resorted to an army of volunteers to gather twenty million 'slips' examples of words in use, with a sentence written out, with details of where it had been found, on an index card. Which were then filed under the word being exemplified, so when he started work on the word, he would go to the filing system, find all the slips for it, and use them as the basis for the entry.

By the 1980s, it was evident that computers could assist and streamline this process enormously. In the UK, this technological development coincided with a commercial one: English Language Teaching was blossoming worldwide, and there was money to be made in dictionaries for the ELT market. The Oxford Advanced Learners Dictionary had, until then, dominated the field, but others were keen to challenge Oxford for a share of the action. In particular, Collins, who found an academic partner in John Sinclair, professor at Birmingham University and already arguing vigorously for corpus methods. Between these two the COBUILD (Collins Birmingham University International Language Database) project was established. It went on to revolutionize dictionary-making, showing that dictionaries could give a more accurate and fuller account of a language if they were based on a corpus.

In 1987, when the COBUILD dictionary was published, Collins were leading the field. By this stage Longman and Chambers also wanted to compete in the EFL dictionaries market, and realized, as did Oxford, that they needed a corpus if they were to keep up. Several Universities were also interested, in particular Lancaster, where Geoff Leech had been an early advocate of corpora in linguistics, for studying topics ranging from grammatical change to stylistics.

Out of this alliance the British National Corpus project was born. The corpus, an unimaginably vast (by the standards of the time) 100 million words, of which 10% were transcribed everyday speech, published in 1994, was the big data of its day. It remains a landmark and model for how a corpus ought to be.

By the 1990s, others were starting to see the potential of corpora, in particular the computational linguists and the technology companies. The computational linguists (also called NLPers 2) asked questions like "if we have a good grammar, shouldn't it account for most of the sentences that we find in a corpus? Shouldn't we be able to treat the grammar as a scientific hypothesis, and, if the hypothesis is good, shouldn't we be able to write a program to find the grammatical structures of the sentences in a corpus?" The technologists were interested both because corpora provided samples of the language that they wanted to be able to handle for purposes of search, or spell-checking, or typesetting, or automatic speech transcription, and because the large corpora and high technological demands that they imposed were a challenge for the technical development of the computers. One innovative project from the early 1990s, HECTOR, was a joint venture between Oxford University Press and DEC, a leading Silicon Valley technology company of the time. DEC wanted to see if their hardware had the potential to index and display results from a very large corpus fast enough, and across enough monitors to meet the lexicographers' needs.[15]

First, we can distinguish two kinds of use of corpora: direct – in the classroom, with students looking at concordances – and indirect, with corpus use by people preparing dictionaries, syllabi, coursebooks and other teaching materials. The success of corpora in indirect use, starting with dictionaries, is clear to see and largely now beyond question. To give an accurate picture of a language, we need evidence of how the language is patterned. To know which phenomena are the common ones, we need language data. For this we need a corpus. Let us turn

[15] www.wikepedia.com

to direct use: ‘corpora in the classroom’. It is here that we might have expected to find explosive growth, in the millions of ELT classrooms worldwide but we have not.

Figure 1

tried to stand up quickly despite his safety	belt	. We thought the bumpy flight must have
form a world-wide network of mountain	belts	far higher (above ocean floor level) and more

should have been awarded a Lonsdale	belt	for his first-half haymaker on Francis . The
the courage to go out and play and not just	belt	the ball up field . ' Saunders , whose penalty
were a relief Like heather flowers His	belt	could not endure the siege - it burst And
Chile-Peru Trench . These steeply-inclined	belts	or zones of earthquake sites are known now
development of ` brown ' areas of the Green	Belt	. UNDERUSED AND SURPLUS PROPERTY IN THE
to reconcile the irreconcilable but green	belt	is one of the most successful all-purpose tools
and put it into his pocket . He found the	belt	at the bottom of the bag . It was a brown
which management regimes are best . Tree	belts	offer many benefits over conventional
Nielsen , the augmented embonpoint , '	belted	' a ' fan ' who tried to give her a ' kiss ' ;
of the width of the contemporary climatic	belts	. But it can hardly be argued that either

Figure 1 shows us how to analyze a basic concordance, for the English word, belt, from the British National Corpus, in a standard ‘key word in context’ form. Concordance of belt (lemmatized) from the British National Corpus. Random sample of twelve lines.

There are several things to note.

- Each line is a fragment of text. It is not a full sentence.

- Each line is from a different text, which is not an authentic experience of language at work.

- Each line has been stripped from its context. If the student wants to lookfurther for more context, to gather more clues to interpretation, they have more work to do.

Thus, each line is not in any sense a self-contained piece of language. Let us now look at the lines one by one.

- The first shows a standard compound for belt, safety belt

- The second line shows a rarer, more specialized compound, possibly familiar only to geologists and similar: mountain belt

- The third line alludes to belts as indications of levels of achievement in somesports, particularly oriental martial arts such as judo and karate, where the practitioner progresses from a beginner's white belt, via a range of colours, to the champion's black belt. This is in fact a report on a rugby match where the journalist is adding colour with a joke from a different sport.

- The fourth is a verbal, informal use: to belt something can mean 'to hit it hard', with the etymology based on schoolchildren being hit with a belt as punishment for bad behavior.

- The fifth is from a poem.

- The sixth, like the second, is geological.

- In the seventh and eighth, the word is part of a technical term, probably only in use in the UK, related to town planning policies, where it is often thought desirable to leave a 'green belt' of agricultural or otherwise 'green' land around towns and cities, and there is UK legislation to make this happen. One of these instances is capitalized, the other is not.

In addition, a further half of the first line is fully capitalized.

- The ninth is (finally) a prototypical and straightforward belt: a narrow piece of leather, cloth etc. that you wear around your waist, for example to keep your clothes in place or for decoration

- The tenth is a technical term from forestry.

- The eleventh is in quotation marks and is another use of the verbal,

informal sense where 'belt' means 'hit'.

- The twelfth is a technical term (which may also be seen as

metaphorical) from climatology.

From a lexicographer's point of view, this is wonderful. In just twelve corpus lines, we have already found nine meanings, most of which need covering in any good dictionary (depending on the size and scope of the dictionary). But from a language learning perspective, it is alarming. Students could not recognize these different meanings based on one meeting with each, and teachers would not encourage them to jump to any conclusion without more evidence. Looking at these lines is not an efficient route to understanding the straightforward English word belt. If we are to succeed in bringing concordances into the classroom, a central challenge is how we do so without scaring the students.[16]

To find out about a word, like English belt, the place to look is the dictionary. The dictionary is designed to give the user the information they need. For learners of English in particular, a number of very high quality monolingual learner dictionaries have evolved to give the student a word's meaning(s), pronunciation, grammar, patterns of use, details of inflection, distributions according to genre such as literary, informal, and regions such as US, UK, Aus. In addition to the traditional book format, they are nowadays

[16] Braun, S., 'Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora', 2007, 79p.

available through internet and mobile phone, with additional searching functionality as well as pronunciation. On the phone, they are always at hand.

So when should a learner turn to the corpus? The short answer is: when the dictionary does not tell us enough. Dictionaries, even online ones, are limited in how much they can say. If entries are too long, information will be hard to find, and users will struggle to find the information they seek. There are various strategies being explored by dictionary-makers to address the issue. Nonetheless, there will always be occasions where the quandary the user wants help with is not covered in the dictionary. There are simply too many cases that different users want help with. This is easier to see when the user is producing language (speaking or writing) rather than receiving it (listening or reading). If the user is writing a chemistry essay and is aware that the appropriate verb for the chemical process is coalesce, but is not sure how the verb should be used, none of the main online dictionaries gives us a chemistry example: it is not such a common word, and gets brief treatment, with examples only in a couple of cases. As we see shortly, if the user looks to corpora they promptly see relevant cases, which can be used, as models for the sentence they are writing.

"But I want to learn English, not find out about corpora" We suspect, the most common reason why students have used corpora in the classroom is because their teacher is a corpus enthusiast. This is not such a good reason, and the student might well say, "but I want to learn English (or French, or Chinese…), not to find out about corpora". How might we overcome students' natural resistance to finding out about something which is, on the face of it, a distraction from the language-learning task?

Here is a possible response. Corpora and dictionaries are both language resources on a spectrum: a spectrum from the raw to the refined. The lexicographer takes the corpus evidence, and then analyses, filters, sorts and selects - and the end product is the dictionary entry.

Students do not know what a corpus is, and maybe do not want to put effort into finding out. But they know what a dictionary is: they are taught how to use them, they regularly do use them, and they quite often even express affection for them. Why not sidestep the question entirely, not introduce corpora, and present corpus evidence as if it was dictionary evidence. Let's disguise the corpus as a dictionary.

Another usage of corpora is related to "word sketch"

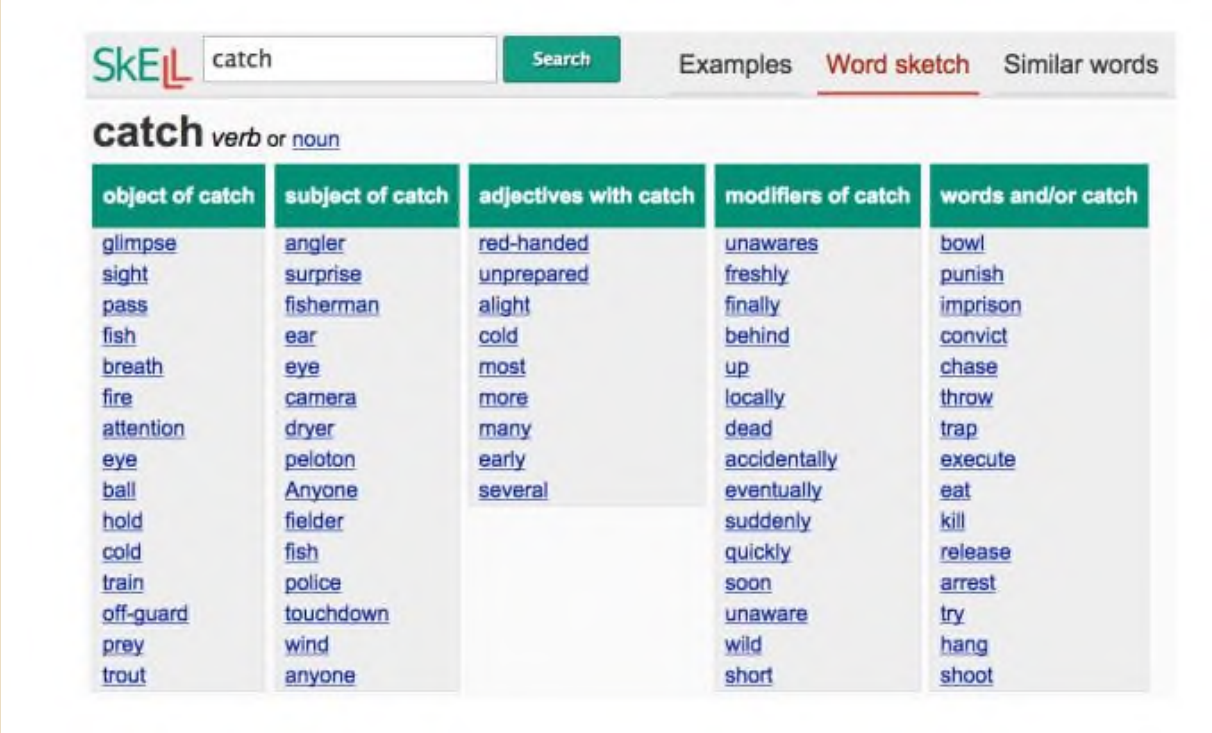


Figure 2: SKELL word sketch for English catch (verb)

The function that gives the Sketch Engine its name is the word sketch. This is a feast of information on the word. For catch (verb) just looking at the first column (objects of the verb) we immediately see a number of meanings, idioms and set phrases. We catch a glimpse of or catch sight of something. Sportsmen and women, in a range of sports, catch passes and balls. Anglers and fisherman (column 2) catch fish, prey and trout. When surprised or shocked, one catches one's breath. Things catch fire. You often want to catch someone's attention. Things sometimes catch your eye. People catch hold of other people

when they don’t want to let them go. We all sometimes catch a cold and catch trains. In the single parsing error in this column, we are often caught off-guard by unexpected events. (Off-guard has been mis-classified as a noun, hence an object, rather than an adverb, hence a modifier.) This is all of great value to learners. The COBUILD dictionary offers thirty meanings of catch, most of which are the verb. It lists the meanings in order of frequency, and, the first two are, in brief,

1. catch a person/animal (capture),

2. catch an object that is moving through the air. These frequent and prototypical meanings have an equivalent in other languages

Beyond this, most of the thirty meanings do not use the "catch" verb e.g. catch a bus: French - prendre le bus, German - den Buserwischen. The word sketch for catch presents learners with clues to the multiple uses of the word, most of which manifest quite differently in their first language. This is valuable for language learning. Moving on to the subject’s column, surprise relates to the expression caught by surprise. Ears and eyes catch things when we hear or see; likewise cameras. Dryers, it turns out, frequently catch fire, and this is often reported, as the cause of a fire, in short local-news reports: we find this is the explanation for dryer being in the list by clicking on the word dryer: we are then shown the ‘examples’ report at Figure 3. In this way the underlying evidence is always available at a mouse-click. In all 40 examples of dryer+catch, it is always caught.

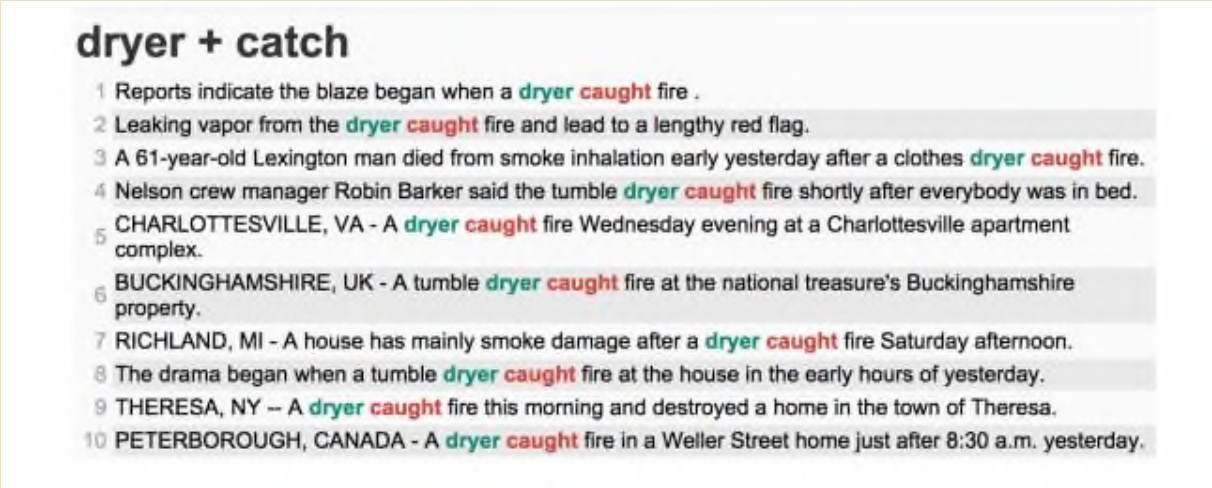


Figure 3: SKELL Examples report for dryer as object of catch

Turning back to the subjects list: peloton is a specialist term from the Tour de France cycling race, referring to the racer who is in the lead and wears the yellow jersey. They often catch the man in front. Fielder and touchdown are also sport terms, both from American football (with another minor mis-analysis: touchdown catch is a compound nominal rather than a subject-verb pair).

Anyone and Police introduce a new meaning of the verb: police catch criminals. Anyone brings to our attention the related pattern Anyone caught [doing X] will be [punished]. The remaining columns tell us more about the police meaning (red-handed, unawares, punish, imprison, convict, chase, trap, execute, kill, release, arrest, try, hang, shoot), the sports/hunting/fishing meaning (locally, wild, bowl, throw, eat) and several other idioms.

For professional and budding lexicographers, the word sketch is a draft dictionary entry. The system has worked its way through the corpus to find all the recurring patterns for the word and has organized them – not by meaning, as a dictionary does: that is too ambitious to do automatically – but at least by grammar, which is some help. As noted above, this is not a report for a beginner learner: all the information that the beginner needs will be provided in a good dictionary. It is for intermediate and advanced learners looking for information they could not find in the dictionary. We find it meets that need well.

On the spectrum from corpus to dictionary, the Examples report is closer to the corpus end. It presents example sentences for the search term. However, in contrast to the concordance for belt in Figure 1:

- each instance is a full sentence, not a series of fragments

- the sentences are chosen as 'good' examples (insofar as this can be done automatically).

The algorithm for choosing good examples is called GDEX (Kilgarriff et al 2008) and works as follows:

- if there is a word sketch for the search term, use that as a starting point and show the best example sentence for each collocation in the word sketch. This will mean that all examples will exemplify a common collocation for the word.

- 'score' each sentence containing the search term, and show the user the highest-scoring sentences. Factors in the scoring algorithm includesentence length: not too long, (so there is too much for the user to struggle through), nor too short (so there is not enough context to be helpful)

- It is a sentence, starting with a capital letter and ending with a full stop.

- Or not too many rare or unrecognized words. This constraint tends to rule out sentences with spelling errors. Mostly comprising common words too many non-letter characters: a large number of numbers, punctuation marks and other non-letter characters tend to indicate a non-standard sentence: not too many capital letters: a large number suggests names and acronyms, which often indicate that the user will need domain-specific knowledge to understand the sentence. Context-free-ness: Ideally, we would like to promote sentences, which stand alone, rather than needing context to be comprehensible. A sentence like "So they decided it was not worth it" is problematic because the reader has no context to help them understand what they, it or it refers to. It was

hard to implement this constraint, and we are unsure if we have been at all successful.

One of the first questions that a corpus linguist will ask about SKELL is: "what corpus do you use?" This is not a question which SKELL immediately presents an answer to. Its users are language learners, not corpus linguists, and we do not expect them to know what a corpus is, let alone to have detailed ideas about what corpus might be good for what task. The learner takes it on trust that a dictionary gives a good account of a language; likewise here. Most corpus tools start by asking the user "which corpus do you want to use?" SKELL views this as a behind-the scenes question that users should not be bothered with.

Behind the scenes, there is of course a full and detailed answer. It was a substantial challenge to prepare a corpus which was big enough – we needed around a billion words to be able to provide good examples for even quite rare collocations - and varied enough, so all text types were covered, and 'clean' enough: without computer-generated junk which looked like English, but was not. Earlier efforts had used web corpora and had included computer-generated material, which was not acceptable as it would give learners bad models. The corpus is described in detail in Baisa and Suchomel.

What comes to similar words, we can state that here corpora can give more effect as well. The third report is similar words: Figure 4 shows similar words for belt (noun). This shows the words that 'share most collocations' with the headword. They are usually words with similar meaning. The word cloud is an attractive way to present the report: the size of each word indicates how similar it is to the target.

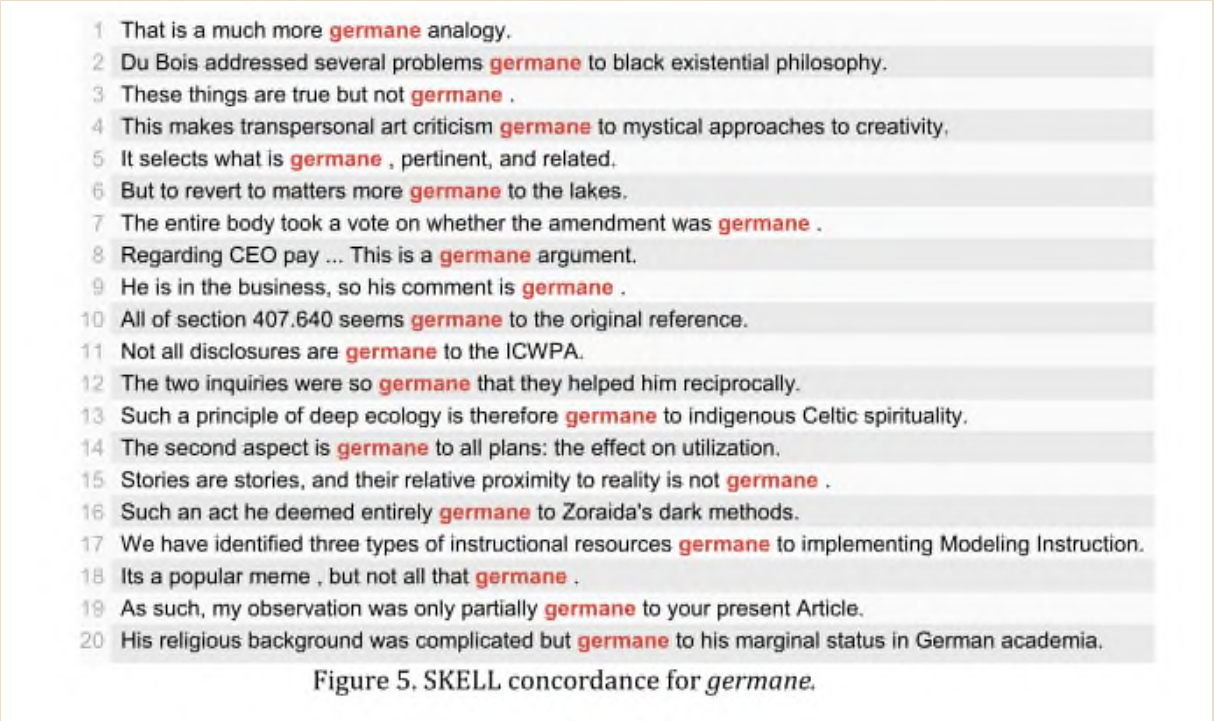
Figure 4: SKELL similar words for English belt (noun).

2.2 Using corpora for teaching lexis and collocation

In this part of the work,we focus on the specific tasks that the learner is asked to do. The role of the tasks is to induct learners into linguistic, cognitive and technical processes.

SKELL is an easy interface to use: the challenge lies in exploiting it, giving learners opportunities to undertake a rich variety of activities, particularly relating to the "semi pre-constructed phrases that we now know are associated with every word". Semi pre-constructed phrases are germane to converting receptive (passive) vocabulary into vocabulary available for productive (active) use. The word germane does not appear in lists for foreign learners. We often meet words in our reading and conversations that pique our interest, but since "authenticity does not automatically entail typicality", it is necessary to look beyond that single first encounter. So let us say a learner wants to know more about this rare bird. The COBUILD dictionary tells us. Something that is germane to a situation or idea is connected with it in an important way; a formal word" and gives us two examples and a grammar code ADJ QUALIT: PRED +to. This is probably not enough information to make it available for confident,

active use. The curious student, having direct access to the data, finds concordances of germane in SKELL, as shown in Figure 5.



1	That is a much more germane analogy.
2	Du Bois addressed several problems germane to black existential philosophy.
3	These things are true but not germane .
4	This makes transpersonal art criticism germane to mystical approaches to creativity.
5	It selects what is germane , pertinent, and related.
6	But to revert to matters more germane to the lakes.
7	The entire body took a vote on whether the amendment was germane .
8	Regarding CEO pay ... This is a germane argument.
9	He is in the business, so his comment is germane .
10	All of section 407.640 seems germane to the original reference.
11	Not all disclosures are germane to the ICWPA.
12	The two inquiries were so germane that they helped him reciprocally.
13	Such a principle of deep ecology is therefore germane to indigenous Celtic spirituality.
14	The second aspect is germane to all plans: the effect on utilization.
15	Stories are stories, and their relative proximity to reality is not germane .
16	Such an act he deemed entirely germane to Zoraida's dark methods.
17	We have identified three types of instructional resources germane to implementing Modeling Instruction.
18	Its a popular meme , but not all that germane .
19	As such, my observation was only partially germane to your present Article.
20	His religious background was complicated but germane to his marginal status in German academia.

Figure 5. SKELL concordance for *germane*.

It is at the next stage that the role of the teacher is critical. The learner cannot be expected to know what linguistic features to look for, nor which ones are germane to the question, at least, not until they have been trained in the requisite metacognitive strategies. Johns recommends the teacher taking on therole of research organizer. This sees the teacher proffering a series of leading questions that the learner can answer from the data:

- Is it typically followed by to? - When it is, what precedes it?
- Does it follow the nouns it qualifies?
- When does it occur at the end of information units?
- Are there any typical nouns or semantic types whose company it keeps?
- Is it used mostly in formal contexts?

The teacher can also create a learning activity using this data: select some sentences and ask if germane could be replaced by any similar words. The

learners arrive not only at semi pre-constructed phrases for germane, but a host of other features. Many mental processes are happening during such work. Thestudent is learning language, about language and about using data. Producing language requires a productive knowledge of vocabulary. Just as we are about to use the word problem, we find that we do not know the collocating verb that expresses the meaning we have in our first language. Here we consider the case where the Czech verb would be spočívat, for which a bilingual dictionary offers ten or so English equivalents. The word sketch provides these fifteen verb collocates with problem as subject: *solve arise face stem occur plaque lie persist confront exist affect beset result concern relate*

Only one, lie, means something similar to the Czech verb. The learner clicks on lie in the word sketch and finds example sentences as in Figure 3, which confirm that they have found the right verb.

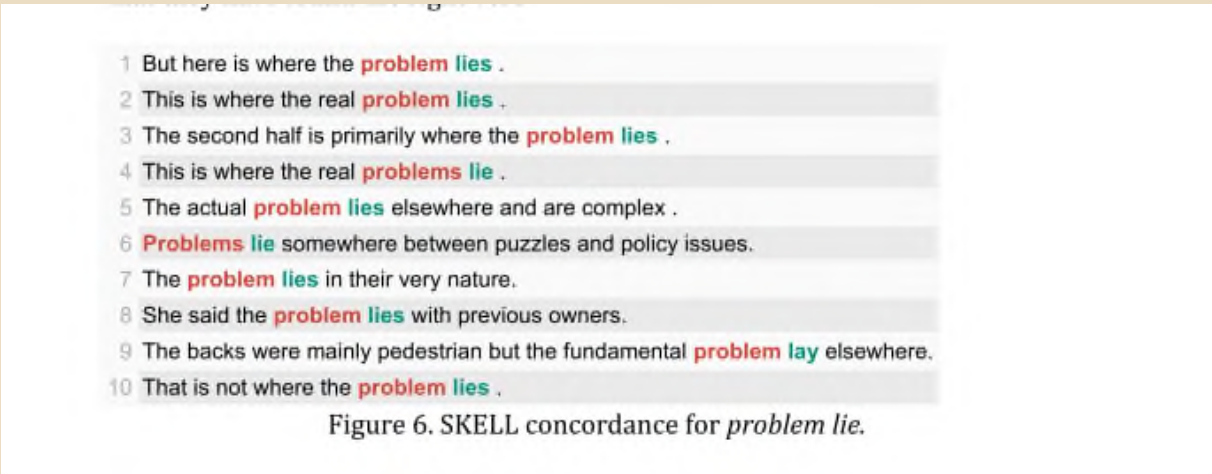


Figure 6. SKELL concordance for *problem lie.*

The productive use of the collocation may require a little more guidance – it is not enough to know problem lie. The teacher may further guide the student with the questions

-What tense is the verb in?

-What article is used?

-Where in the sentence is the expression?

The curious student might even wonder what other abstractions lie? An answer to this question could lead to seeing this use of lie as a pattern, not just an idiosyncratic collocation. Such an observation triggers a new appreciation of how words are combined to create text that sounds idiomatic and native-speaker like. The nouns given in the word sketch for lie, in the 'subject' column, are answer fault snow strength body future island blame difficulty land danger village loyalty interest problem. By asking the students to divide this list into concrete and abstract nouns, the curious students have answered their question. To develop their understanding, they click on the abstract nouns to see what follows lie in these sentences. The solution to this problem lies not in legislation but in effective screening of potential surrogate mothers.

To observe the patterned nature of this, we type lie not in into the Examples search field and get Figure 7.

Figure 7. SKELL concordance for *lie not in.*

Many abstract nouns are the subject of lie, and not predicts a following but. These concordances reveal a pattern the X abstract lies not in Y, but in … which the learner may use in a sentence they write.

It may be objected that no one has time for these shenanigans in class, as there is a strict syllabus to be followed. This makes it all the more important for teachers to appreciate how many learning experiences are happening at the same time. All being well, the teachers have inducted their students into the basics, and the students can practice discovery tasks for homework.

Teachers set tasks for students so that students become familiar with the many kinds of question that SKELL data can answer. Without this guidance, students will find little of interest. But once they have a framework, they become able to ask interesting and pertinent questions, and answer them.[17]

Learners may pursue language study for only a short period of their university career, but once the corpus is constructed, students may be sufficiently motivated to consult it and add to it when needed.

The large corpus can then be used in the following ways:

1. To produce lists of subject area words and terms for study

2. To view word sketches, which give a one-page view of the collocations and grammatical structures in which a word or term participates.

3. To view the words and terms in context, using concordancing

4. To link back to the original texts on the web.

2.3 Corpora as a tool of student autonomy and independence

The idea of learner autonomy is not new, but it has been widely referred to in the field of ELT only over the last decade. The notion of learner autonomy was first developed out of practice by teacher-researchers at the Centre de Recherches et d'ApplicationsPédagogiquesenLangues (CRAPEL), University of Nancy, France, in the early 1970s. In the interests of widening access to education and promoting lifelong learning, CRAPEL began to offer adults the opportunity to learn a foreign language in a resources centre, free from teacher direction. CRAPEL put in place various kinds of support measures, including learner counselling and 'training' to assist in the 'autonomization' process—the

[17]Campoy, M., Gea-valor, M. and Belles-Fortuno, B., Corpus-based Approaches to English Language Teaching. London: Continuum, 2010, 39p.

development of learners' abilities to work more effectively in a self-directed fashion. It soon became clear that participants have the full capacity to take charge of decision-making in all the areas normally determined by an institution, teacher, or textbook. This views learner autonomy as a capacity and willingness to act independently, "the ability to take charge of their learning" , for example learning by themselves, choosing and studying material satisfying their needs, understanding whether they have some problems with English or how well they are prepared for the test, reading books not assigned by a teacher and so on. It doesn't mean that they work completely independently, without a teacher or completely alone. Moreover, it doesn't mean that all learners have it, but they all have a capacity to become autonomous. In order to foster learner autonomy, it is necessary to develop a sense of responsibility and encourage learners themselves to make decisions about their learning. What is more, it's possible to do at any age.

Corpora is normally associated with noticing and language awareness. Gabrielatos draws attention to the fact that corpora can also be used by teachers following more traditional methodologies, such as PPP (Presentation—Practice—Production), as well as task-based approaches. He argues this because corpora can be used to look up examples of the language the teacher (or the students) wish to focus on, for instance in a PPP lesson, instead of using invented sentences which contain the target language (e.g. should and must), the teacher could obtain sample sentences containing should and must from a corpus, then teach the lesson in the typical fashion. However, we should not forget that although corpora can be used to design a more traditional style of lesson like PPP, researchers normally see corpora use as helping to create a less teacher-centered classroom atmosphere.Different investigations suggest that using corpora promotes students' autonomy, because students can take responsibility for their own language learning, looking for language features in the different existing language data bases with the teachers' guidance. Instead of

relying on the teacher for information, corpora can be used by learners to find things out for themselves. Students' autonomy is not a threat to teachers. Gabrielatos argues that teachers' roles in corpus-based classrooms won't be any less important: their role will only be different This is not to say that the teacher's role is diminished; rather, it is enriched and diversified. The teacher becomes less a provider of input and facts about language and more a facilitator and consultant, or, at the learner-centered end, a co-researcher.Corpora use also empowers non-native teachers because it helps them to be independent. For instance, they can find suitable examples of 'real life' language, which makes them, feel more confident about the language they are presenting to the students.Nowadays in a globalized world it is impossible to conceive of education without technology and all the advantages that it has given to the teaching- learning process in general, and to the language teaching-learning process in particular. It is difficult not to realize that computers are used to inform teachers in developing classroom activities as well as to facilitate their job. The use of corpora is a new tool that provides teachers with authentic data about language structure and also promotes student autonomy because they can explore a determined corpus and do their own research about language features. Unfortunately, using corpora is not an easy task, especially for those teachers from institutions whose aim is not just English teaching (typical public school with limitations), and thereforeit is possible that even when English teachers are informed of this new technology, they may continue to base their teaching on textbooks with dubious language content.Schools require investing in computer equipment as well as in software to analyze corpora in order to introduce this new technology and teachers and students can make use of its advantages. Despite the facteconomical problems (lack of computers in the school) would be enough reason for teachers not to use corpora to supplement their teaching course. They can use corpora in a basic way at home with the help of one computer and internet using the free downloadable software to analyze a corpus.

They can identify examples of authentic language use for their regular classes no matter the methodology they use to teach.

A corpus is a collection of texts or utterances (made easier with the advent of the computer and electronic databases), assembled for the purpose of studyinglinguistic structures and patterns. Current corpora can store millions of words. A concordancing software can analyze the linguistic features of the texts. A concordancer can, for example, show the frequency of the occurrence of words and/or their collocates.Some online corpora come with their own concordancing software, thus making it easier for the user. Some, however, require you to process it through your own concordancers. In such cases, one could invest in a good concordancing software like Wordsmith. Alternatively, there are also freeware concordance programs like those on Laurence Anthony's website.For the purpose of this blogpost, we are going to reference Brigham Young University's corpus of global web-based English, which makes use of the British National Corpus, is free and easy to access (although frequent use would require you to register and set up an account, all of which is free of charge), and comes with their own concordancing software.

Imagine a learner coming across the newspaper headline:

'Government to boost economy'.

Knowing every word in that headline apart from the word 'boost', the learner proceeds to guess its meaning with some help from the news story.

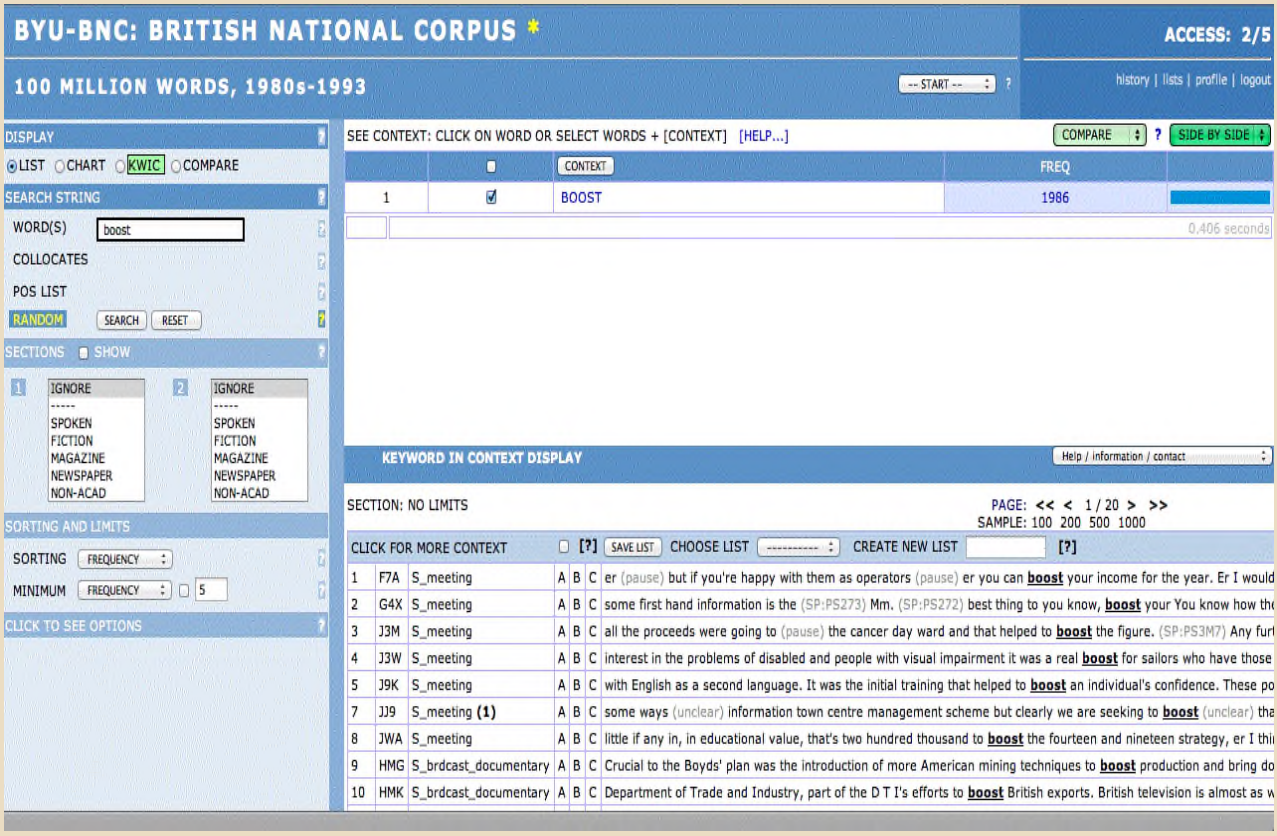
But how can the learner find out more about this word.

He/She could of course look it up in the dictionary, which would tell him/her that 'boost' could mean 'to lift', 'to raise' or 'to increase'. It might even give the learner a couple of examples, like 'to boost prices' and 'to boost the horsepower of the car' (courtesy of dictionary.com).

The learner now goes away knowing a little more about the word ‘boost’. But would he/she feel confident enough to produce it in a sentence?

Could the learner, for example, say ‘I must boost my hand when I ask a question in class’, or ‘Her pocket money boosted when she reached 12’, or even ‘I want to boost my English vocabulary’?

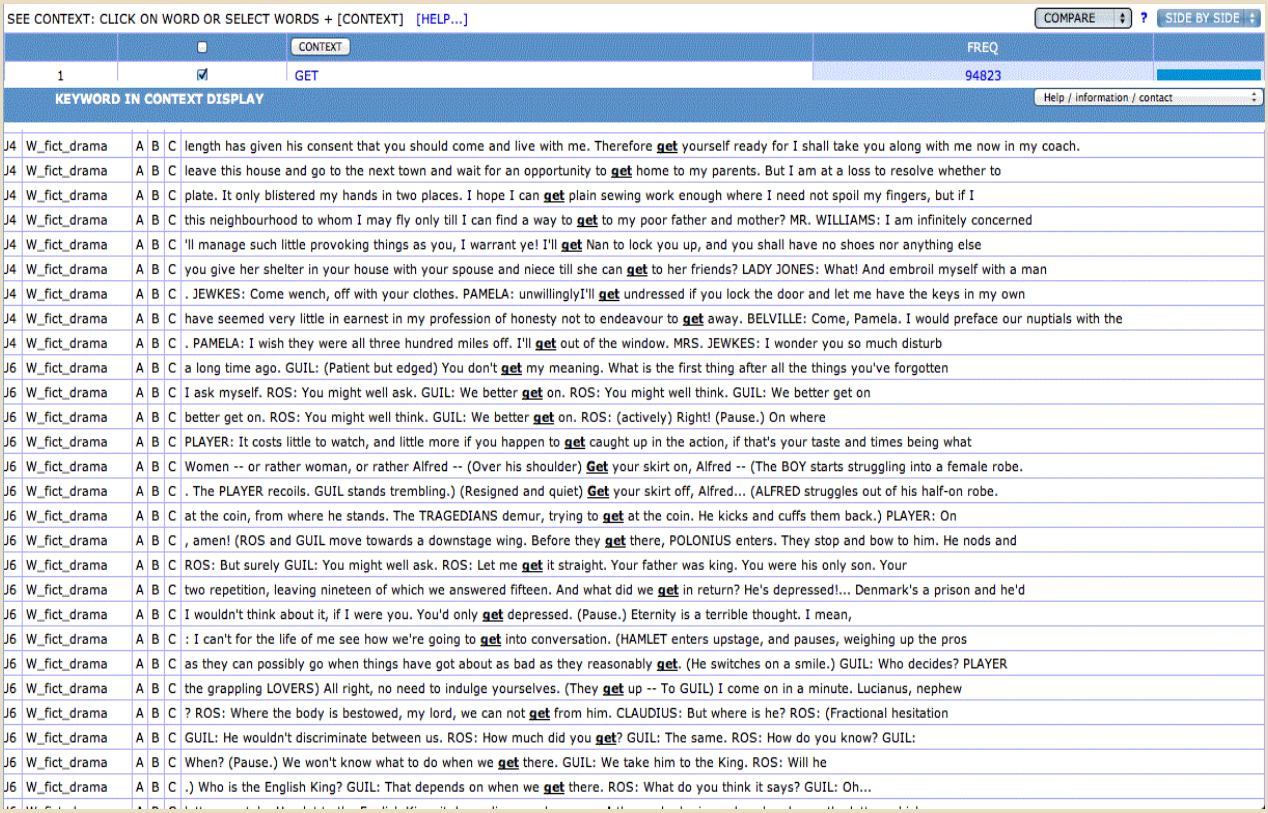
How could the learner know whether those are sentences that are commonly used or not, without always seeking the help of an English teacher?



In the case of *‘boost’*, one might find that the majority of example sentences come from newspaper and tabloid, followed by those from academic writing in the fields of law, politics and education. This alone could inform us as to when and where one might come across the word ‘boost’.

We know that words do not exist on their own, and that their meanings often depend on the surrounding words that accompany it.

Nevertheless, where does a phrase start and end? Getting students to practice 'chunking' a text cannot only help them understand the lexis used, but also enable them to identify patterns of usage. Let's try to identify some of the chunks with the word *'get'*.



SEE CONTEXT: CLICK ON WORD OR SELECT WORDS + [CONTEXT] [HELP...] COMPARE ? SIDE BY SIDE

		CONTEXT	FREQ
1	☑	GET	94823

KEYWORD IN CONTEXT DISPLAY — Help / information / contact

J4	W_fict_drama	A	B	C	length has given his consent that you should come and live with me. Therefore **get** yourself ready for I shall take you along with me now in my coach.
J4	W_fict_drama	A	B	C	leave this house and go to the next town and wait for an opportunity to **get** home to my parents. But I am at a loss to resolve whether to
J4	W_fict_drama	A	B	C	plate. It only blistered my hands in two places. I hope I can **get** plain sewing work enough where I need not spoil my fingers, but if I
J4	W_fict_drama	A	B	C	this neighbourhood to whom I may fly only till I can find a way to **get** to my poor father and mother? MR. WILLIAMS: I am infinitely concerned
J4	W_fict_drama	A	B	C	'll manage such little provoking things as you, I warrant ye! I'll **get** Nan to lock you up, and you shall have no shoes nor anything else
J4	W_fict_drama	A	B	C	you give her shelter in your house with your spouse and niece till she can **get** to her friends? LADY JONES: What! And embroil myself with a man
J4	W_fict_drama	A	B	C	. JEWKES: Come wench, off with your clothes. PAMELA: unwillinglyI'll **get** undressed if you lock the door and let me have the keys in my own
J4	W_fict_drama	A	B	C	have seemed very little in earnest in my profession of honesty not to endeavour to **get** away. BELVILLE: Come, Pamela. I would preface our nuptials with the
J4	W_fict_drama	A	B	C	. PAMELA: I wish they were all three hundred miles off. I'll **get** out of the window. MRS. JEWKES: I wonder you so much disturb
J6	W_fict_drama	A	B	C	a long time ago. GUIL: (Patient but edged) You don't **get** my meaning. What is the first thing after all the things you've forgotten
J6	W_fict_drama	A	B	C	I ask myself. ROS: You might well ask. GUIL: We better **get** on. ROS: You might well think. GUIL: We better get on
J6	W_fict_drama	A	B	C	better get on. ROS: You might well think. GUIL: We better **get** on. ROS: (actively) Right! (Pause.) On where
J6	W_fict_drama	A	B	C	PLAYER: It costs little to watch, and little more if you happen to **get** caught up in the action, if that's your taste and times being what
J6	W_fict_drama	A	B	C	Women -- or rather woman, or rather Alfred -- (Over his shoulder) **Get** your skirt on, Alfred -- (The BOY starts struggling into a female robe.
J6	W_fict_drama	A	B	C	. The PLAYER recoils. GUIL stands trembling.) (Resigned and quiet) **Get** your skirt off, Alfred... (ALFRED struggles out of his half-on robe.
J6	W_fict_drama	A	B	C	at the coin, from where he stands. The TRAGEDIANS demur, trying to **get** at the coin. He kicks and cuffs them back.) PLAYER: On
J6	W_fict_drama	A	B	C	, amen! (ROS and GUIL move towards a downstage wing. Before they **get** there, POLONIUS enters. They stop and bow to him. He nods and
J6	W_fict_drama	A	B	C	ROS: But surely GUIL: You might well ask. ROS: Let me **get** it straight. Your father was king. You were his only son. Your
J6	W_fict_drama	A	B	C	two repetition, leaving nineteen of which we answered fifteen. And what did we **get** in return? He's depressed!... Denmark's a prison and he'd
J6	W_fict_drama	A	B	C	I wouldn't think about it, if I were you. You'd only **get** depressed. (Pause.) Eternity is a terrible thought. I mean,
J6	W_fict_drama	A	B	C	: I can't for the life of me see how we're going to **get** into conversation. (HAMLET enters upstage, and pauses, weighing up the pros
J6	W_fict_drama	A	B	C	as they can possibly go when things have got about as bad as they reasonably **get**. (He switches on a smile.) GUIL: Who decides? PLAYER
J6	W_fict_drama	A	B	C	the grappling LOVERS) All right, no need to indulge yourselves. (They **get** up -- To GUIL) I come on in a minute. Lucianus, nephew
J6	W_fict_drama	A	B	C	? ROS: Where the body is bestowed, my lord, we can not **get** from him. CLAUDIUS: But where is he? ROS: (Fractional hesitation
J6	W_fict_drama	A	B	C	GUIL: He wouldn't discriminate between us. ROS: How much did you **get**? GUIL: The same. ROS: How do you know? GUIL:
J6	W_fict_drama	A	B	C	When? (Pause.) We won't know what to do when we **get** there. GUIL: We take him to the King. ROS: Will he
J6	W_fict_drama	A	B	C	.) Who is the English King? GUIL: That depends on when we **get** there. ROS: What do you think it says? GUIL: Oh...

Did you find *'get ready'*, *'get someone to do something'*, *'get my meaning'*, *'to find a way to get to someone'*, *'get undressed'*, *'Let me get it straight'*, and *'get depressed'*, among many others?

We can have students categorize the different meanings of a word (hyponyms).

Can you categorize the word *'get'* in the example sentences into its different meanings?

How about categorizing the different meanings of *'board'* that you can find in the above examples?

Perhaps one category could have something to do with transport, e.g. *'they can board of the famous San Francisco trams', 'the jumbo would have had up to 450 passengers and 12 crew on board', '...before they are allowed to board their plane',* etc.

Another category uses board to mean *'a flat piece of wood'* such as *'...scrawled on a board'* and *'...pinned on the notice board'*.

Yet another category is concerned with food/meals, as in *'40 per adult half board'*.

But one of the most common use of *'board'* among these example sentences seem to be one with the meaning of *'a group of people either administering a company/organization or on a committee/council'*, e.g. *'the Legal Aid Board'*, *'the Irish Tourist Board'*, *'the examination board'*, etc.

We can have students compare two near-synonyms.

Our learners often come to me with questions like *"What's the difference between 'listen' and 'hear'?"* and *"What's the difference between 'trouble' and 'problem'?"*

We would sometimes be able to give a satisfactory answer because I have either read up on these differences in a teacher's book or blogpost, or have come across this question before in my teaching career and have thought extensively about the answer.However, there are often no hard and fast rules to language, especially when it comes to lexis, and we could not always vouch for the validity of my answers (or those from a teacher's book, for that matter).Furthermore, by taking on the responsibility as 'knowledge giver', I seem to be robbing my learners of a chance to discover the answers for themselves.Using the corpus, I'm able to pull up example sentences of words that my learners have problems with, and allow them to compare and contrast the near-synonyms for themselves.Here's a real-life example that I experienced in my language classroom.

A group of advanced students wanted to know the difference between the words *'squeamish'* and *'queasy'*. Admittedly, they were not words that you would find in an English teacher's handbook, nor ones that you are asked about by most students.Pulling up example sentences of *'squeamish'*,

Conclusion

Corpora, plural term of a 'corpus', refer to electronic authentic language databases that can be available via internet or as software installed in desktops. Language corpora can be either collections of written or spoken texts; for example, collections of written texts can be extract from newspapers, business letters, popular fictions, books, or magazines, published or unpublished school essays and etc. Collections of spoken texts can be any recorded formal or informal conversations, radio shows, weather broadcasts or even business meetings and etc.The use of corpora in language teaching and learning has been more indirect than direct. This is perhaps because direct use of corpora in language pedagogy is restricted by a number of factors including, for example, the level and experience of learners, time constraints, curricular requirements, knowledge and skills required of teachers for corpus analysis and result interpretation, and the access to resources such as computers, and appropriate software tools and corpora, or a combination of these.

A Corpus (plural Corpora) is a large collection of written texts, which are used, in computational linguistics for analysis of the way language is used. They are most often analyzed using a concordancer.

Due to insufficient learning time and inefficient word searching tools, lexical learning has always been one of the main language learning problems that learners pointed out. The advent of technology is about to present a different view toward language learning and teaching; several studies have shown positive learning outcomes by engaging students in activities of decision making and information retrieving. Integration of corpora into vocabulary classrooms not only provides learners faster searching tools and better quality of contexts that traditional dictionaries are not likely to achieve but enhance their learning motivation.

Language corpora can be used by anyone who is engaged in language learning, teaching, or research; language learners or even native speakers may find it useful to assist academic writing or lexical knowledge (Qiao, 1995); teachers can utilize the authentic collections of data as classroom materials for ESL, EFL, or EAP (English for Academic Purpose) learners; language researchers or linguists often use corpora as language sources to analyze certain aspects of a certain language.Usually users of corpora use the searching tool, the concordance, to look for vast number of authentic language contexts analyzed from corpora; this feature provides users not only better quality of examples but also more exposures to an unfamiliar word.The British National Corpus is a 100-million-word text corpus of samples of written and spoken English from a wide range of sources. The corpus covers British English of the late twentieth century from a wide variety of genres with the intention that it be a representative sample of spoken and written British English of that time.Of the two parts to the 10-million word spoken corpus, one is a demographic part, containing transcriptions of spontaneous natural conversations made by members of the public and the other a context-governed part, containing transcriptions of recordings made at specific types of meetings and events. All the original recordings transcribed for inclusion in the BNC have been deposited at the National Sound Archives of the British Library.

There are some types of corpora. A corpus can be one or more of the following:monolingual; multilingual; general texts; texts on a specific subject or genre e.g. scientific papers, Shakespeare plays only, children's essays, etc.; texts from a specific varieties of English, e.g. American English or British English, etc.

Corpora are generally searched and analyzed using computers which are able to search and compare millions of text strings in virtually no time. However, computer analysis does sometimes have drawbacks. For example, take these two sentences:

1. Time flies like an arrow.

2. Fruit flies like a banana.

Whilst a human can easily distinguish between the two uses of the words, flies and like a computer does not yet find this possible. To get around this, corpora are often tagged or annotated. Typically, this would involve human operators giving parts of speech tags to words before they are processed and compared by the computer,thus:

1. Time [noun] flies [verb] like [adverb] an [determiner] arrow [noun].

2. Fruit [adjective] flies [noun] like [verb] a [determiner] banana [noun].

This allows, for example, a concordancer to analyze all uses of like as a verb as oppose to like as an adverb.

Analysis of a corpus will bring to light certain ways of language use within that group. For example, it may well show that scientific papers use the passive voice far more often than newspapers do or that certain words are only used among certain groups of speakers.

Uzbekistan is implementing language corpora in use turn by turn, systematically; currently we can state that very few examples of classes using corpora are conducted. But language corpora can assist teachers to conduct better and easier classes if they are used appropriately.

References:

1. Aijmer, K. Corpora and Language Teaching. Amsterdam: John Benjamins, 2009.
2. Allan, Q. 'Enhancing the language awareness of Hong Kong teachers through corpus data'. Journal of Technology and Teacher Education 7/1, 1999.
3. Aston, G. (ed.) Learning with Corpora. Houston, TX: Athelstan, 2001.
4. Baker, M., 'Corpus linguistics and translation studies: implications and applications'. In M. Baker, G. Francis and E. Tognini-Bonelli (eds.) Text and Technology: in Honour of John Sinclair. Amsterdam: Benjamins,1993
5. Ball, F., 'Using corpora in language testing'. Research Notes, 2001
6. Bernardini, S., Competence, Capacity, Corpora: A Study in Corpus-aided Language Learning. Bologna, 2000
7. Biber, D., Conrad, S. and Reppen, R., Corpus Linguistics: Investigating Language Structure and Use. Cambridge: Cambridge University Press,1998
8. Braun, S.,'Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora', 2007
9. Burnard, L. and McEnery, A. (eds.), Rethinking Language Pedagogy from a Corpus Perspective. New York: Peter Lang, 2000
10. CampoyM., Gea-valor, M. and Belles-Fortuno, B., Corpus-based Approaches to English Language Teaching. London: Continuum, 2010
11. Johansson S., 'Contrastive linguistics and corpora'. In S. Granger, J. Lerot and S. Petch-Tyson (eds.) Corpus-Based Approaches to Contrastive Linguistics and Translation Studies. Amsterdam, 2003
12. Keck, C.'Corpus linguistics and language teaching research: bridging the gap'. Language Teaching Research 8(1), 2004

13. Kennedy, G., An Introduction to Corpus Linguistics. London: Longman, 1998
14. Mindt, D.'English corpus linguistics and the foreign language teaching syllabus' in J. Thomas and M. Short (eds.) Using Corpora for Language Research, London: Longman, 1996
15. Mishan, F. Designing Authenticity into Language Learning Materials. Chicago: Chicago University Press, 2005
16. Römer, U. 'Corpora and language teaching'. In A. Lüdeling and M. Kyto (eds.) Corpus Linguistics: An International Handbook,Berlin: Mouton de Gruyter, 2008
17. www.ziyonet.uz
18. www.languagecorpora.com
19. www.wikepedia.com