Brezina, V. forthcoming. "Google Scholar as a linguistic tool: new possibilities in EAP" C.Gkitsaki, and R. Baldauf (eds.), *The Future of Applied Linguistics: Local and Global Perspectives*. Newcastle upon Tyne: Cambridge Scholars Publishers.

CHAPTER XXXX

GOOGLE SCHOLAR AS A LINGUISTIC TOOL: NEW POSSIBILITIES IN EAP

Abstract: The present paper introduces *Google Scholar* (*GS*) as a linguistic tool available to researchers, practitioners and students in English for academic purposes (EAP). It demonstrates how the academic search engine can be used for searching not only for academic *content*, but also for the *form* of academic expression. In particular, the paper discusses the possibilities of employing *Google Scholar* as a tool for exploring collocations in academic language. It shows how the large *GS virtual corpus* of written academic English can be effectively used in EAP research as well as in creation of corpus-informed teaching materials.

1. Introduction

One of the theoretical positions generally accepted in corpus linguistics is the conviction that language cannot be conceptualised as merely a composite of grammatical rules on the one hand and lexis (or vocabulary) on the other. Instead, as the evidence from large language corpora suggests, there exist important patterns in language between lexis and grammar—prefabricated chunks available to users that make their production natural and fluent (Barlow, 2011; Römer, 2009; Hunston & Francis, 2000). These patterns have come to be called *collocations* (sometimes also referred to as *lexical bundles* or *multi-word patterns*) (Biber, 2009, 2006; Sinclair, 1991a; Firth, 1957: 194ff). In the present paper, the term *collocation* will be used in this general sense to mean any frequent co-occurrence of two or more words in text.

The aim of the paper is to introduce *Google Scholar* (*GS*) as a linguistic tool, which can be used to explore collocational patterns in a virtual corpus of academic writing and thus help (novice) writers produce naturally sounding texts.

2. Background

When John Sinclair, one of the pioneers of corpus linguistics, discussed the workings of language, he postulated the *idiom principle.* The idiom principle states that "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments" (Sinclair, 1991a:110). There is an overwhelming evidence from language corpora suggesting that speakers (or writers) indeed do not choose words completely freely (and randomly), restricted only by the subject matter, social considerations and the genre (cf. Barnbrook, 2009; Hunston, 2002; Sinclair, 1991a). On the contrary, each single selection of a word has a large impact on the selection of the following words.

In order to produce a naturally sounding piece of text (written or spoken) one needs to acquire and master certain preferred ways of expression (or set of collocations) appropriate for a given situation. This is especially true in academic writing, which is a set genre with relatively strict conventions (Gotti, 2009; Swales, 2004, 1990). Acquiring appropriate academic collocations—the building blocks of academic language—can be therefore viewed as an important step on the way to becoming an expert writer.

Exploring academic collocations, however, goes beyond the limits of any existing dictionary (although learner dictionaries such as the *Cobuild* or *Oxford Advanced Learner's Dictionary* pay particular attention to multi-word expressions). The most suitable tool for investigating the preferred ways of expressing things in academic writing as well as for the development of teaching materials is a large corpus of academic texts comprising hundreds of millions of words. At present, however, no such corpus is available. The current corpora used in English for academic purposes (EAP) are all relatively small—not larger than several million running words (Krishnamurthy & Kosem, 2007; Flowerdew, 2002). Moreover, most of the written language corpora in EAP were created for a particular research and are not generally available.

One of the solutions to this problem would be to build a new corpus of academic writing from the scratch. However, this solution would be extremely time consuming and would require extensive human and financial resources (not to speak of the additional problems with copyright). Although it is not generally recognised, there exists one potential corpus resource for academic writing which is freely available to anyone with internet access. This resource is the *Google Scholar* search engine, which can access large academic text databases (*GS virtual corpus*). As I will argue in this article (and illustrate with examples of teaching materials), *GS* can be effectively used for both EAP research and teaching practice.

The aim of this paper is thus to show that the *GS virtual corpus* can be searched not only for academic *content* (which is the intended use of the search engine) but also for the *form* of academic expression. In particular, the paper

discusses the possibilities of employing *Google Scholar* as a linguistic tool for creating teaching materials, which can help students to identify useful collocations in academic writing.

2.1 Corpora in EAP

Generally speaking, there are three main virtues of a corpus: *large size*, *representativeness* and *availability*. Although smaller corpora (comprising several million running words) usually used in EAP research (Krishnamurthy & Kosem, 2007; Hunston, 2002: 198-204) can be very useful in identifying some frequent lexico-grammatical patterns, their size may not be sufficient for exploring even fairly common collocations. Sinclair (2004: 189-190) demonstrates this with the collocational phrase *fit into place*. We need a corpus of at least 200 million running words to get several examples of this phrase. Similarly, if we want to search for a frequent academic expression such as *have a large impact on*, it is not enough to go through the 600 pages (about 200,000 words) of the *Applied linguistics* journal published in one year. In volume 30, it does not appear even once.

Moreover, for analysis of internal variation within collocational patterns, we need even larger corpora. Do we write *have a large* or *great* or *major* or *profound* or *significant* or *substantial impact?* Are there any contextual preferences for one rather than another of these qualifications? Are there any other qualifications that can fill in the adjective slot in the structure? To be able to answer these questions, we need to have an access to a large corpus of hundreds of million or even several billion running words.

Representativeness is another keyword which has been often discussed in relation to language corpora (e.g. Nevalainen, 2001; Tognini-Bonelli, 2001: 57ff; Biber, 1993; Atkins et al., 1992). It is important to realise that no matter how large, a corpus will always be only a sample of language. It can never be a collection of *all* texts, spoken and written, ever produced. The question therefore is: How well does the sample (i.e. a particular corpus) represent language? Can we make generalisations based on this sample about the whole of language or a particular genre, i.e. academic writing in our case?

Moreover, in academic writing, we can find different disciplinary conventions (Hyland, 1999). Some of these conventions are immediately apparent from a cursory look at two journal articles from two different fields (such as Applied linguistics and Physics), some are more subtle and require a careful analysis. The requirement for a good academic writing corpus will therefore ultimately be to represent a range of academic disciplines and their particular conventions.

Finally, the third aspect of a good corpus which, however, is not often discussed in literature, is its availability. Only if a corpus of academic writing is available to a larger research community and also to EAP practitioners, can it have some major positive impact on the development of the discipline. Researchers will be able to replicate previous research and base new research on comparable data. Practitioners will be able to create teaching materials and use the corpus in the classroom.

As we shall see in the next section, *Google Scholar virtual corpus* satisfies all three criteria discussed above. It is a large corpus, which represents a variety of academic disciplines and is available to anyone with internet access.

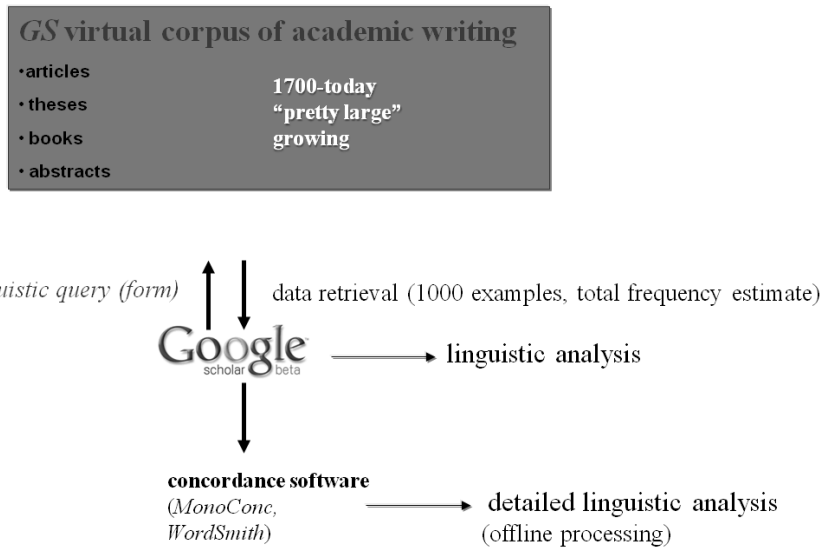
2. 2 *Google Scholar* in EAP

Although *GS* has received a lot of attention from numerous scholars, especially in computer and information studies (e.g. Howland et al., 2009; Howland et al., 2009; Jacsó, 2008), to the best of my knowledge, it has not yet been systematically discussed as a linguistic (rather than academic) tool. When using *GS* for linguistic purposes, the first thing we need to realise is that *GS* itself is *not* a corpus, but an academic search engine, which provides access to a large index of academic texts such as research articles, theses, books and abstracts ("About Google scholar," 2010). We can call this index *GS virtual corpus of academic writing* (see Fig. XXXX-1).

Although the oldest texts *GS* can access are some early prints from the 18th century, the bulk of the material indexed comes from the current academic production. It is not possible to estimate exactly the size of the corpus. However, we can be relatively certain that the size of *GS virtual corpus* is in the range of dozens (if not hundreds) of billions of running words (Lewandowski & Mayr, 2006). For a comparison, Lew (2009) estimates that the size of the textual resources on the whole World Wide Web was approximately 5 trillion tokens in 2005. It is supposed (Zhang et al., 2008) that between 2005 and 2010 the World Wide Web doubled its size.

Moreover, *GS* index is updated on a regular basis to include the most recent academic texts. Through *GS* we therefore have access to a large corpus of academic texts, which is constantly growing. In this respect, *GS virtual corpus* comes close to Sinclair's idea of *monitor corpus* (Sinclair 1991a, pp. 24 - 26) "because of its capacity to hold a 'state of the language'" (p. 26).

Fig. XXXX-1. *Google Scholar* virtual corpus



The procedure of performing linguistic searches in *GS* is relatively simple. If we want to carry out a basic linguistic search we need to type our query into the query box accessible through the standard *GS* interface (www.scholar.google.com). We can also make use of the *Advanced Scholar search*, which enables us to search for articles written by a particular author and/or published in a particular journal and/or published within a particular date range. In addition, we can limit the search to broadly defined disciplinary fields such as biology, life sciences and environmental sciences; social sciences, arts and humanities; physics, astronomy and planetary science etc. Nevertheless, before we start with linguistic analyses, it is important to change the number of displayed results in *Scholar Preferences* from 10 (default) to 100 (maximum) in order to be able to inspect more examples at once.

GS search returns a maximum of 1000 results (with a maximum of 100 results per page) and indicates the approximate estimate of the total number of documents satisfying the query. As has been, however, pointed out in the literature and numerous internet blogs (Kilgarriff, 2007; Liberman, 2005; Nunberg, 2005), the estimates are imprecise and often contradictory.

The results of the *GS* search can be inspected visually in a web browser and the main collocational patterns can be observed (a basic type of linguistic analysis). The following (Fig. XXXX-2) is an example of one of the results of a search for the definite article "the":

Fig. XXXX-2. Structure of *GS* search results

Why are children in **the** same family so different from one another?	TITLE
R Plomin, D Daniels - Behavioral and Brain Sciences, 2010 - journals.cambridge.org	SOURCE
The theme of **the** target article is that environmental differences between children in **the** same family (called "nonshared environment") represent **the** major source of environmental variance for personality, psychopathology, and cognitive abilities. One example of **the** evidence**...**	FULL- TEXT LINES
Cited by 560 - Related articles—All 4 versions	OTHER DETAILS

As can be seen from the example above, all occurrences of the search term in the *GS* results are displayed in bold type. The three full text lines offer us a context of 30 to 50 words, in which the search term appears. This is similar to the usual stretch of context that standard concordances offer for a KWIC (key word in context) in corpus linguistics.

If we, however, want to carry out a more detailed analysis, we need to download the results (copy them into a text file) and analyse them using a standard concordance software package such as *MonoConc* or *WordSmith*. For a more detailed discussion and examples of these two types of analysis see Section 3 below.

The last point that needs mentioning in this section is the formulation of linguistic queries (i.e. expressions we can type into *GS* search box). Although *GS* has not been designed as a linguistic search engine, it can be successfully used for linguistic searches. Linguistic searches can be defined as specific types of searches, which target not the *content* but the *form* of linguistic expressions. For the purpose of linguistic searches, we often need to employ *GS* search operators (see Table XXXX-1).

Table XXXX-1. *GS* operators

Operator	Explanation	Example
Simple searches		
Double quotation marks ""	Exact phrase search	"in fact"
Minus sign -	Exclude the word	-the
Plus sign +	Sic! (Search for the given form)	+lingvistik +linguist
Asterisk *	Any single word	"as * points out"
Double full stop ..	Number range search	"as * 1990..2010 pointed out"
Complex searches		
OR	OR	his OR her
AND	AND (implied)	his her, his AND her
Parentheses ()	Embedded searches	"(points OR pointed OR pointing) out"
		"(fulfill OR fulfills OR fulfilled OR fulfilling) * (obligation OR obligations)"

The most useful operators are double quotation marks (""), which enable us to search for an exact phrase, an asterisk (*) which replaces any single word, and double full stop (..) which is used to search for a number range. We can also formulate complex searches with the aid of Boolean operators (AND and OR) and parentheses.

3. The Study

In this section, two examples of teaching materials (see Appendix A and Appendix B) which invite students to use *GS* virtual corpus to identify collocations in academic writing will be discussed. In addition, some initial observations from the teaching practice using these materials will be reported. The materials were developed for academic writing classes at the University of Auckland, New Zealand. All students in these classes were non-native speakers of English with varying degrees of English proficiency (pre-intermediate to upper-intermediate). The following are the main aims with which in mind the teaching materials were created:

- To introduce students to a powerful language learning tool (GS virtual corpus)
- To introduce students to *GS* search syntax
- To get students to identify useful collocations in different parts of research articles

There are two types of linguistic analysis which can be employed in the teaching practice in three different ways. As was mentioned above, *GS* virtual corpus can be subjected to: 1) basic analysis (i.e. visual inspection of the results returned by *GS* without further quantification) which can be used for identification of the basic tendencies in the data and 2) complex offline analysis using standard concordance software (*MonoConc, WordSmith*, etc.) which yields more detailed (and quantifiable) results.

For the standard classroom purposes, the basic type of analysis is often sufficient, especially given the time limitations in academic writing courses. In such a case, students need to learn only the elementary principles of *GS* search syntax and understand the idea behind collocational patterns and their variation. After this, they are ready to carry out linguistic searches in a (for most of them) familiar *GS* search environment. The major advantage of this approach is its simplicity. Students can inspect the data (results of the searches) using a standard web browser without the need to install any specialised software. The only prerequisite is access to a computer with internet connection. Students can therefore also do the basic analysis easily outside of the classroom.

For more sophisticated analyses, we need to introduce students to concordance software. This requires extra time and can present a challenge for the instructor (Flowerdew, 2009; Hunston, 2002). The disadvantage of most concordance packages is that they have been developed primarily for corpus researchers and students in EAP courses might find them difficult to use. However, if students master the basic concordancing techniques, this will give them a very powerful tool for analysing large amounts of language.

The middle ground is that the students are instructed to do the basic analysis while the instructor provides the results of a complex analysis as part of the feedback. Since both types of analysis are based on the same data set, the conclusions will be similar. The instructor, however, will be in a position to provide more details about the variation in individual collocations and supply also frequency information, i.e. which of the collocations are more frequent than others.

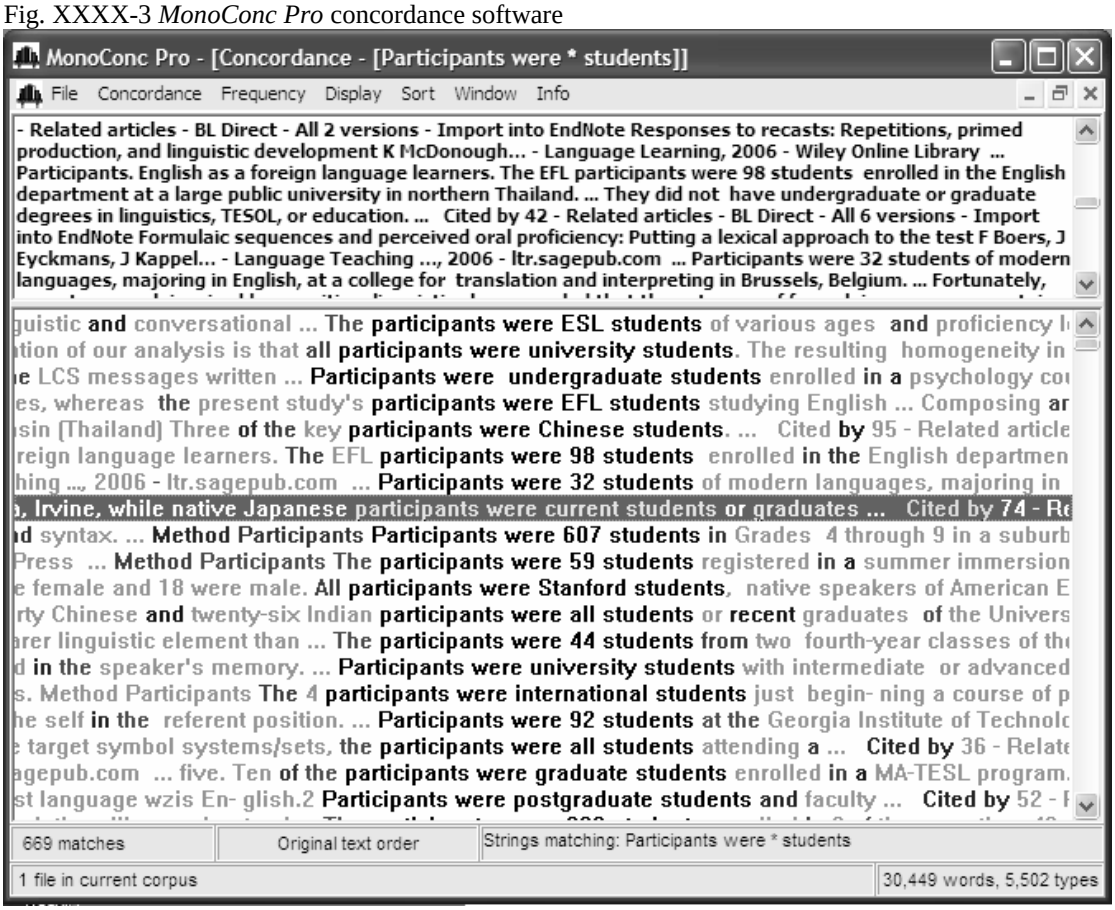
3.1 Teaching materials

This section demonstrates the possibilities of linguistic analyses using *GS* worksheets in the Appendices and the *GS* search engine. The objective of the first worksheet (see Appendix A) is to introduce students to the language of the *Method section* of written academic reports (e.g. theses, research articles etc.).

The first exercise (*Population and sample*) does not focus on the exploration of any fixed expressions (in the strict lexicographical sense) in academic English. Instead, it invites students to use the two basic operators of *GS* search syntax, quotation marks ("") and asterisk (*), in order to investigate what kind of information about the participants is usually included in research reports.

Visual inspection of the results returned by *GS* will offer a variety of qualifications in the structure *participants were * students*. However, we can gain a more detailed insight into this structure if we download the first 1000 results returned by *GS* (which is the maximum of results we can extract from a single search) and subject them to standard corpus linguistic analysis. For the present analysis, concordance software *MonoConc Pro* (see Fig. XXXX-3) was used (cf. Barlow, 2002).

Fig. XXXX-3 *MonoConc Pro* concordance software



After the errors and duplicates were removed, 672 examples of the structure *participants were * students* remained for further analysis. As we can see from Table XXXX-2, the slot was most often filled by the number of participants (258 instances), followed by the level of students (150 instances), the type of institution they belong to (98 instances), participants' nationality (16) and their characteristics (11). Finally, the slot was also occupied by the adverbial modifiers (9 instances) *mostly* and *mainly*.

Table XXXX-2. Analysis of the *participants were * students* structure

Type of complementation of the structure	Frequency	Details (individual frequencies)
<NUMBER>	258	
<LEVEL>	150	undergraduate (94), graduate (47), postgraduate (6), doctoral (3)
<INSTITUTION>	98	university (74), college (24)
<NATIONALITY>	16	international (10), German (2), Japanese (2), foreign (2)
<PARTICIPANT CHARACTERISTICS>	11	female (7), young (2), male (2)
<ADVERB>	9	mostly (5), mainly (4)

Those interested also in the methodological aspects of academic research (i.e. not only the purely linguistic ones) can also take a closer look at the number of participants who were recruited for the studies which appear in *GS* results. It is important to note that in my search, I limited the results to the studies which were related to language research by including the context word *linguistics* in the search box. From the analysis, we can see that there was a large diversity in the participant numbers ranging from 4 to 6,313. However, the median was 60, which is an expected number in social science research.

Part b) of Exercise 1 looks at further modifications of the structure *participants were * * students* with two slots in the middle. Similarly, the rest of the exercise [c)—e)] further explores the possibilities of the structure with different national groups of participants (*Malay*, *Chinese, Finish*) included. The national groups were selected so as to represent the student population in the academic writing courses. Students could thus more readily relate to these examples.

The aim of Exercise 2 (*Instruments & procedure*) is to explore collocational patterns NOUN + VERB (passive) + PREPOSITION typical of the Method section. Here, the discussion will be limited to the first structure *the questionnaire was <past participle> <preposition>*.The other two phrases in Exercise 2 are analogous.

Table XXXX-3 offers results of a detailed analysis of collocational patterns using *MonoConc Pro.* It is based on the analysis of 779 examples downloaded from *GS* after errors and duplicates were deleted.

Table XXXX-3. Ten most frequent collocations of the structure *the questionnaire was...*

Collocations	**Freq.**	**Example**
designed to (elicit, assess, explore)	32	*The questionnaire **was designed to** elicit students' attitudes to....*
administered to	27	*The questionnaire **was administered to** 18 South Korean secondary school EFL teachers.*
given to	20	*The questionnaire **was given to** the subjects*
administered in	17	*The questionnaire **was administered in** Chinese.* *The questionnaire **was administered in** the last session of the term.*
divided into	15	*The questionnaire **was divided into** three parts of 18 items each*
distributed to	15	*The questionnaire **was distributed to** 1,731 tenth graders*
based on	12	*The questionnaire **was based on** (Nass and Brave, 2005) and (Mutschler et al., 2007).*
translated into	11	*The questionnaire **was translated into** Russian and Polish by native speakers (all linguists).*
piloted with	10	*The questionnaire **was piloted with** 30 learners.*
sent out	9	*The final version of the questionnaire **was sent out to** 200 high-school teachers.* *The questionnaire **was sent out** and returned by post.*

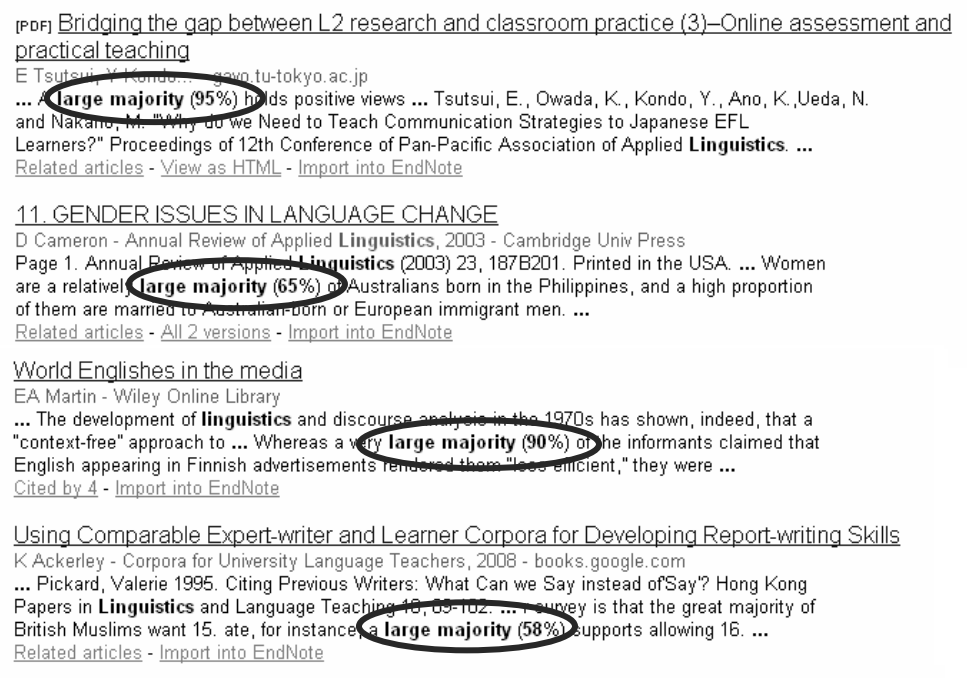
We can see that the most frequent collocation in the data is *the questionnaire was designed to...* stating the purpose of the instrument, followed by the structures *the questionnaire was administered to..., the questionnaire was given to..., the questionnaire was distributed to...* and *the questionnaire was sent out..* which all describe the procedure of the questionnaire distribution. The other phrases are used to speak about the details of the questionnaire and its design (*the questionnaire was divided into/based on/ translated into/ piloted with...)* and the circumstances of the questionnaire distribution (*the questionnaire was administered in...)*.

The second teaching material (see Appendix B), which engages students with the *Results section* of the research reports, introduces a new *GS* operator *double full stop* (..) which can be used to search for a number range. For example, if we type *1..100* into the *GS* search box and press enter we ask *GS* to search for any number between 1 and 100.

The first exercise is designed to draw students' attention to generalising expressions of proportion such *vast majority, slight majority, almost half*, etc. and to the way they are used in the context of academic writing. Students

are asked to search for the percentages which often appear in parentheses after a particular expression of proportion. When doing the exercise students should note the variable use of these expressions. Let us take the expression *large majority* as an example. The variable use of this phrase is apparent already from the first page of the *GS* results (see Fig. XXXX-4).

Fig. XXXX-4. *GS* results: "large majority 1..100"



If we subject the results returned by *GS* to a further analysis, we can be more precise about the percentages (although this is not the primary aim of Exercise 1). However, it is probably interesting to note that there is a slight difference between how the expression is used in Natural sciences on the one hand and Humanities and social sciences on the other (see Table XXXX-4). Although the percentage range is similar in both fields, authors in Natural sciences on average use the expression *large majority* to mean a higher percentage (83%) than authors in Social sciences and humanities (78.6%).

Table XXXX-4. The use of the expression *large majority*

Disciplinary field	**Examples analysed**	**Percentage range**	**Mean**
Natural sciences	556	51-99%;	83%
Humanities & soc. Sciences	468	53-99%	78.6%

Exercise 2 in Appendix B asks students to identify verbs, which the authors of academic texts use to refer to the location of results such as the verb *show* in the expression *Table 1 shows*. In fact, we are looking for specific NOUN + VERB collocations, in which the NOUN slot is occupied by the word *Figure* or *Table*. The exercise also invites students to inspect more pages with *GS* results in order to identify verbs other than the relatively obvious *show* and *present.*

The visual inspection of several pages of the *GS* search results offers the following candidates: *Figure 1..50 illustrates/compares/summarises* and *Table 1..50 gives/lists/summarises/reports*. From a detailed analysis, we can obtain the following collocations (see Table XXXX-5 below).

Table XXXX-5. Verbs used in the location of results structures

	Verbs	Frequency	Example
Figure 1..50	shows	296	Figure 1 **shows** the structure of the present EL cell.
	illustrates	48	Figure 2 **illustrates** the results obtained with the first seven constructs.
	presents	12	Figure 2 **presents** the mortality experience of a population of actively employed male workers...
	depicts	10	Figure 1 **depicts** the model and describes some of the issues to be resolved at each stage.
	compares	9	Figure 5 **compares** the content of these two views.
	displays	9	Figure 1 **displays** the distribution of blocking activities...
	gives	8	Figure 1 **gives** the modified hierarchy.
	demonstrates	7	Figure 1 **demonstrates** the sharp decline ...
	summarises	6	Figure 2 **summarizes** the model of investor preferences...
	represents	6	Figure 4 **represents** the dependence of agglutinability on enzyme concentration.
	provides	4	Figure 2 **provides** an example of a defect profile...
	indicates	2	Figure 3 **indicates** the actions followed by the tag array control...
	plots	2	Figure 8 **plots** the lifetime of networks gathering data from sources...
Table 1..50	shows	139	Table 2 **shows** the relative distribution of cases...
	gives	36	Table 2 **gives** the percentage of each of these four groups...
	summarises	32	Table 5 **summarizes** the results.
	lists	31	Table 1 **lists** the observed IR line positions...
	presents	26	Table 1 **presents** the mean values of cholesterol content in HDL separated by the three methods stated.
	displays	6	Table 2 **displays** the correlation matrix for the state variables.
	demonstrates	5	Table 2 **demonstrates** the effect of nucleotides and related compounds on cAMP binding.
	indicates	5	Table 3 **indicates** the percent of patients with either daily heartburn or only monthly heartburn
	contains	5	Table 5 **contains** the correlational results of this re-analysis
	represents	5	Table 1 **represents** the allocation of profiles among raters.
	compares	4	Table 1 **compares** the estimates of length obtained for each sample.
	reports	3	Table 1 **reports** the results of two specifications of the regression equation.
	describes	2	Table 1 **describes** the relationship of SSS ratings to the performance of Ss.

We can see that in the *GS* sample, there are 13 verbs which collocate with *Figure* and the same number of verbs that collocate with *Table* in the location of results phrase. Although most of the verbs are used in both *Figure* and *Table* expressions, there are also notable differences. First of all, the frequency order of the collocates is different. Despite the fact that *show* is the most popular verb on both lists, it is followed by *illustrate, present* and *depict* on the *Figure* list, while in the *Table* list, the next verb is *give* followed by *summarise* and *list.* Moreover, most of the verbs which are unique collocates with *Figure* have visual-graphic connotations (*illustrate, depict, plot*). On the other hand, verbs uniquely associated with *Table* (*list, contain, report, describe*) are verbs of verbal presentation.

3.2 Classroom experience

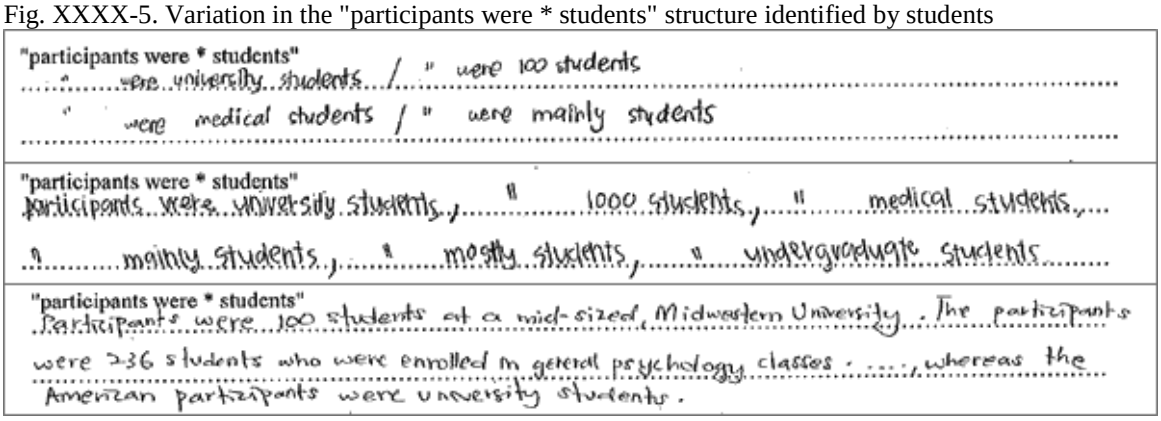
This section offers a brief discussion of the initial experience with using *GS* materials in the academic writing courses mentioned above. In these courses, students were asked to perform various types of basic analyses of *GS* data as part of their homework. The instructor then provided the results of complex analyses as part of the feedback.

Overall, students found the materials engaging. Their enthusiasm for *GS* activities can partly be attributed to the novelty of the exercises. Long-term effectiveness of these activities, therefore, remains to be investigated.

One of the most positive outcomes of the introduction of *GS* in the classes was the fact that students were able to learn quickly how to formulate complex linguistic queries in *GS* and apply the technique of collocation searchers to new situations. This is probably not surprising considering the fact that the majority of them were familiar with the *GS* online environment.

The following are examples of the patterns students were able to identify through the basic analysis of *GS* data (visual inspection) in Exercise 1 from the first teaching material (see Appendix A):

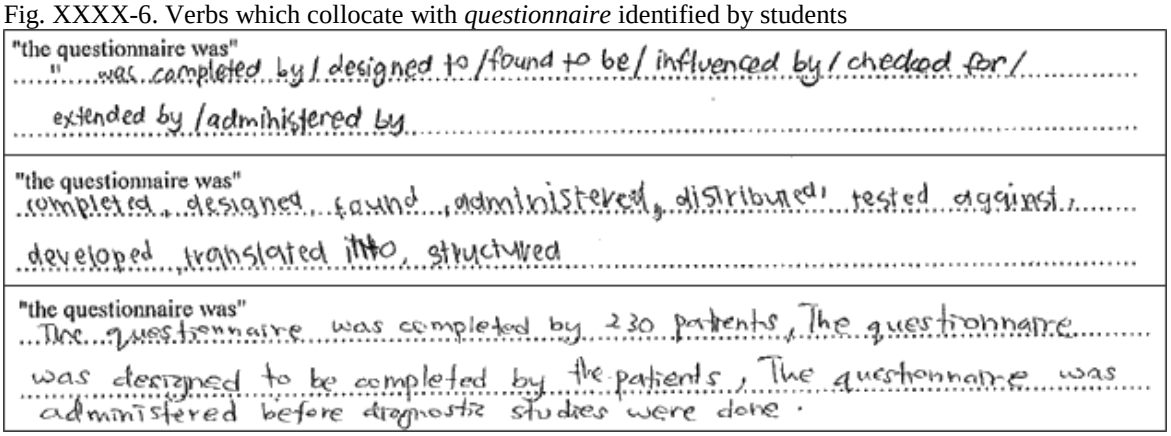
Fig. XXXX-5. Variation in the "participants were * students" structure identified by students



As we can see from Fig. XXXX-5, students noticed that the empty slot (marked by an asterisk) in the phrase *participants were * students* can be filled with different types of qualifications (*university, medical, undergraduate* etc.), the quantification (the number of participants) and an approximator (adverbs such as *mainly* or *mostly*).

Further examples show the collocates of the structure *the questionnaire was* (Appendix A, Exercise 2) which students identified:

Fig. XXXX-6. Verbs which collocate with *questionnaire* identified by students



As can be seen from Fig. XXXX-6, students identified a number of verbs which collocate with the noun *questionnaire* in this passive structure, such as *complete, design, administer* etc. Students also noticed a range of prepositions (*by, for, into, before*) which follow these verbs.

4. Conclusion

The present paper sought to show that apart from the individual corpora of academic writing which are usually small and difficult to get access to, EAP researchers and practitioners have a powerful corpus tool at their disposal, which is easy to access and also relatively easy to use. This tool is the *GS virtual corpus of academic writing*. Being a large corpus, it provides us with a valuable insight into the conventions of academic writing and allows us to explore useful collocational patterns in academic language.

Although using *GS* for linguistic purposes has many advantages, it also presents us with some challenges and limitations. The major limitation to stress here is the fact that the user does not have a full control over the corpus. *GS*

virtual corpus is not a corpus in the strict traditional sense—a carefully selected sample of language. Instead, it is a constantly growing index (database) of academic texts, which are searchable through a simple web browser interface. Besides, in the search results we can only access first 1000 examples which are sorted by a not very transparent principle of relevance.

Despite these limitations, linguists, teachers and students who use *GS* to search for *forms* (rather than *content*) now have a new linguistic tool at their disposal. *GS* makes it possible to explore academic collocations, which cannot be discovered in existing dictionaries, grammar books or small corpora for that matter. What is more, it has the potential to change our perspective on academic language as it enables us to engage ourselves with the subtleties of linguistic variation which often manifest themselves through a variety of collocational patterns (and their internal variation). Sinclair believed that "a new understanding of the nature and structure of language will shortly be available as a result of the examination of large collections of text"(Sinclair, 1991b, p. 489). With *GS virtual corpus* we may stand at the beginning of such a promising journey in EAP.

References

About Google scholar. (2010). Retrieved 14/6, 2010, from http://scholar.google.com/intl/en/scholar/about.html

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and linguistic computing, 7*(1), 153-164.

Barlow, M. (2002). *MonoConc Pro 2.0*. Houston: Athelstan Publications.

Barlow, M. (2011). Corpus linguistics and theoretical linguistics. *International journal of corpus linguistics, 16*(1), 3-44.

Barnbrook, G. (2009). Sinclair on collocation. In R. Moon (Ed.), *Words, grammar and text: Revisiting the work of John Sinclair* (pp. 23-38). Amsterdam: John Benjamins.

Bergh, G., Seppänen, A., & Trotta, J. (1998). Language corpora and the Internet: A joint linguistic resource. In A. Renouf (Ed.), *Explorations in corpus linguistics* (pp. 41-54). Amsterdam: Rodopi.

Biber, D. (1993). Representativeness in corpus design. *Literary and linguist computing, 8*(4), 243-257.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International journal of corpus linguistics, 14*(3), 275-311.

Firth, J. R. (1957). *Papers in linguistics, 1934-1951*. London: Oxford University Press.

Fletcher, W. H. (2007). Concordancing the web: promise and problems, tools and techniques. In M. Hundt, N. Nesselhauf & C. Biewer (Eds.), *Corpus linguistics and the Web* (pp. 7-24). Amsterdam: Rodopi.

Flowerdew, L. (2002). Corpus-based analyses in EAP. In J. Flowerdew (Ed.), *Academic discourse* (pp. 95-114). Harlow: Longman.

Flowerdew, L. (2009). Applying corpus linguistics to pedagogy: A critical evaluation. *International journal of corpus linguistics, 14*, 393-417.

Gotti, M. (Ed.). (2009). *Commonality and individuality in academic discourse*. Bern: Peter Lang.

Howland, J. L., Howell, S., Wright, T. C., & Dickson, C. (2009). Google scholar and the continuing education literature. *The journal of continuing higher education, 57*(1), 35-39.

Howland, J. L., Wright, T. C., Boughan, R. A., & Roberts, B. C. (2009). How scholarly is Google scholar? A comparison to library databases. *College & Research Libraries, 70*(3), 227-234.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hunston, S., & Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Hyland, K. (1999). Academic attribution: citation and the construction of disciplinary knowledge. *Applied linguistics, 20*(3), 341-367.

Jacsó, P. (2008). Google Scholar revisited. *Online information review, 32*(1), 102.

Kilgarriff, A. (2007). Googleology is bad science. *Computational linguistics, 33*(1), 147-151.

Krishnamurthy, R., & Kosem, I. (2007). Issues in creating a corpus for EAP pedagogy and research. *Journal of English for academic purposes, 6*(4), 356-373.

Lew, R. (2009). The Web as corpus versus traditional corpora: Their relative utility for linguists and language learners. In P. Baker (Ed.), *Contemporary corpus linguistics*. London: Continuum.

Lewandowski, D., & Mayr, P. (2006). Exploring the academic invisible web. *Library Hi Tech, 24*(4), 529-539.

Liberman, M. (2005). Questioning reality. Retrieved 31.12., 2010, from http://itre.cis.upenn.edu/~myl/languagelog/archives/001837.html

Lüdeling, A., Evert, S., & Baroni, M. (2007). Using Web data for linguistic purposes. In M. Hundt, N. Nesselhauf & C. Biewer (Eds.), *Corpus linguistics and the Web* (pp. 7-24). Amsterdam: Rodopi.

Nevalainen, S. (2001). Corpora: Corpus representativeness - a query. Retrieved 31.12., 2010, from http://www.hit.uib.no/corpora/2000-3/0144.html

Nunberg, G. (2005). When things don't add up. Retrieved 31.12., 2010, from http://itre.cis.upenn.edu/%7Emyl/languagelog/archives/001834.html

Römer, U. (2009). The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual review of cognitive linguistics, 7*, 141-163.

Sinclair, J. (1991a). *Corpus, concordance, collocations*. Oxford: Oxford University Press.

Sinclair, J. (1991b). Shared knowledge. In J. E. Alatis (Ed.), *Linguistics and language pedagogy: the state of the art* (Vol. Georgetown University round table). Washington, D. C.: Georgetown University Press.

Sinclair, J. (2004). *Trust the Text: Language, corpus and discourse*. London: Routedge.

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Swales, J. M. (2004). *Research genres: Explorations and applications*. Cambridge: Cambridge University Press.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: John Benjamins.

Zhang, G.-Q., Yang, Q.-F., Cheng, S.-Q., & Zhou, T. (2008). Evolution of the Internet and its cores. *New journal of physics, 10*, 1-11.

APPENDIX A

A teaching material developed by the author for academic writing courses for undergraduate students.

Google Scholar: useful language in the METHOD section

Use *Google Scholar* (scholar.google.com) to identify common collocations (i.e. bits of useful language) in the *Method section* of research reports.

- Search for a whole phrase using quotation marks ("")
- Use an asterisk (*) to replace any single word.

Write down TWO or THREE examples of useful collocations from each search.
E.g.: *The **participants were ESL students** of various ages; **Participants were undergraduate students** enrolled in...*

❶ Population and sample

a) "participants were * students"
……
……..

b) "participants were * * students"
……
……..

c) "subjects were * Malay"
……
……..

d) "subjects were * Chinese"
……
……..

e) "subjects were * Finnish"
……
……..

❷ Instruments & procedure

a) "the questionnaire was"
……
……..

b) "the test was"
……
……..

c) "the following instruments were"
……
……..

APPENDIX B

A teaching material developed by the author for academic writing courses for undergraduate students.

Google Scholar: useful language in the RESULTS section

Use *Google Scholar* (scholar.google.com) to identify common collocations (i.e. bits of useful language) in the *Results section* of research reports.
Search for a whole phrase using quotation marks ("")

- Use an asterisk (*) to replace any single word.
- Use double full stop (..) to indicate a number range, e.g. 1..100

❶ Search for the percentages that often occur with the **Expressions of proportion** in the Table below. Use the suggested search phrases in quotation marks.

E.g. A search for "vast majority 1..100" returns the following results:

[PDF] Prevalence of serum antibody to staphylococcal enterotoxin F among Wisconsin residents: implications for toxic-shock syndrome
JM Vergeront, SJ Stolz, BA Crass, DB Nelson, JP ... - The Journal of infectious ..., 1983 - JSTOR
... Although recent surveillance data suggest that TSS occurring in males may be more common than previously appreciated [16], **a vast majority (97%)** of the TSS cases reported to the Wisconsin Division of Health [6] and 96% of the cases reported to the Centers for Disease ...
Cited by 113 - Related articles - All 4 versions - Import into EndNote

Relationship between the seed rain and the establishment of vegetation in two areas abandoned after peat harvesting
V Salonen - Ecography, 1987 - interscience.wiley.com
... 7.2 12.8 * 3685.2 2.8 6.0 2.0 0.4 8.4 46.0 0.4 0.8 0.4 2.0 10.0 119.3 0.4 not estimated IIOI.AKC lie I-COLOGY I(>:.1 (1987) 173 Page 4: new habitat (van Hulst 1978). Of the seeds dispersing to the study areas **a vast majority (99%)** possessed some kind of a transportation device. ...
Cited by 60 - Related articles - All 5 versions - Import into EndNote

Expressions of proportion	Search for..	Range of percentages in RR
vast majority	"vast majority 1..100"	99%—70%
large majority		
substantial majority		
small majority		
slight majority		
less than half	"less than half 1..100"	
almost half	"almost half 1..100"	
nearly one third	"nearly one third 1..100"	

❷ Search for the verbs commonly used in LOCATION OF RESULTS such as the verb *show* in the expression "***Table 1 shows"***. Try to identify verbs other than *show* and *present.*

Search for...	Verbs used for location of the results
"Figure 1..50 * the"	
"Table 1..50 * the"	