"Value added teaching": corpus-based methods for LSP teaching

TISLID'10

James Wilson, Serge Sharoff, Paul Stephenson

Centre for Translation Studies, University of Leeds

ABSTRACT

Corpora have assumed a strong foothold in learning and teaching during the last decade and their application in LSP teaching is particularly beneficial, given the lack of conventional teaching materials in several domains. In this paper, we present a corpus-based approach to teaching business Russian at the University of Leeds, describing how we have enhanced our existing corpus-based tools to facilitate vocabulary acquisition and register recognition and differentiation.

Keywords: LSP, corpus, frequency lists, genre classification

1 INTRODUCTION

Corpora have assumed a strong foothold in learning and teaching during the last decade and their application in Language for specific purposes (hereafter, LSP) teaching is particularly beneficial, given the lack of conventional teaching materials in several domains. In this paper, we present a corpus-based approach to teaching business Russian at the University of Leeds, describing how we have enhanced our existing corpus-based tools to facilitate vocabulary acquisition and register recognition and differentiation.

The paper was motivated by students' concerns over the lack of relevant materials for learning business Russian with English-language commentary. Students complained that there was not a course book containing materials on Russian business terminology, conventions in formal etiquette and style or sample official documents of various kinds (CVs, covering letters, letters of complaint, etc) – documents which students are expected to produce in class tests and for homework assignments (and, of course, in the "real world"). We describe how corpora can be used to compensate for the lack of such materials and how a corpus-based approach is particularly relevant to LSP teaching.

The most important areas that need to be addressed on the business Russian module are: (1) vocabulary acquisition and (2) register recognition and differentiation. The module is taken at Level 2 (in students’ third year at university) and, although by this stage students have studied Russian for at least two years and have spent nine months in Russia, it is their first “real” experience of formal writing. The lexicon of business Russian is different to that of other genres that students are already familiar with, and many key words and phrases are not included in standard dictionaries. Furthermore, even at the (upper-)intermediate level, students are not fully competent in recognising and successfully differentiating between different varieties of written Russian. Our aim is therefore is to use corpora to facilitate vocabulary acquisition and to enhance students’ ability to recognise and use formal expressions appropriately. To achieve these aims, we have developed our corpus-based tools to:

(1) make it easy for users to compile frequency lists;
(2) automatically extract collocations;
(3) automatically classify text genres and domains;
(4) rank concordance lines according to their difficulty;
(5) allow users to collect and annotate their own corpora.

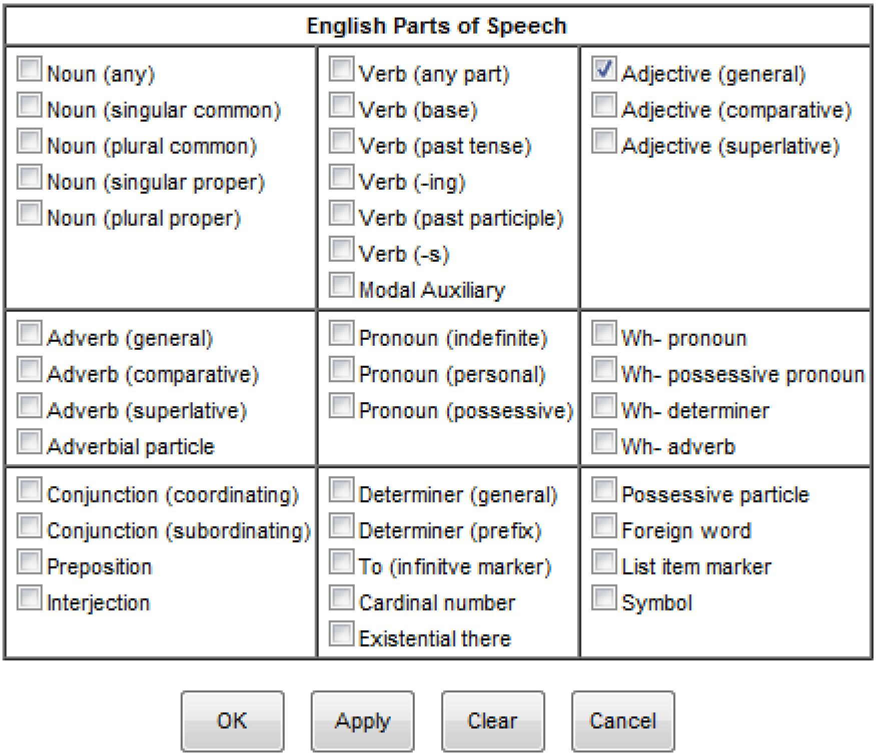
2 Simplifying and enhancing corpora for teaching

In most cases, corpora are not designed for language learners and must be adapted for language teaching. As large collections of texts corpora, especially those that are unannotated, are of little practical value to language learners. What, for example, can you do with the output of a corpus search? Other than for reference a list of concordance lines is of little use to teachers and their students. A corpus that is parsed and tagged is more useful; language learners can make more elaborate corpus quires, searching for, say, a noun in a particular case, a verb in a particular tense or a multi-word expression. However, users need in many cases to be familiar with regular expressions, as most interfaces require grammatical information to be entered as a string code. Assume that a language learner wants to see which adjectives are commonly used in the expression *to make an impression* between the words *an* and *impression*. In many interfaces they would to enter the following string code (CQP syntax):

```
[lemma="make"] [pos="DT"] [pos="JJ"] [word="impression"]
```

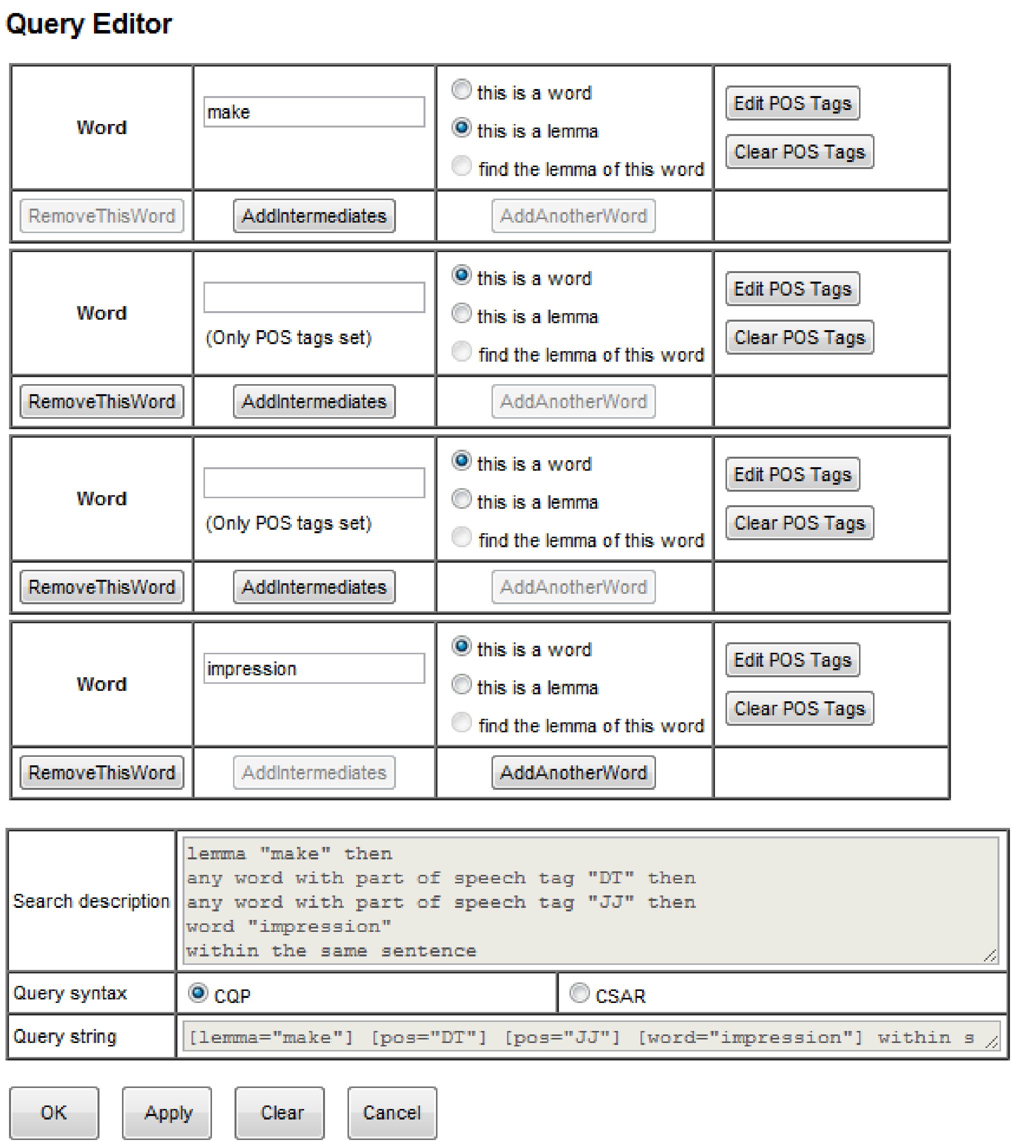
The string tells the regular expression processor to look for any form of *make*, a determiner (DT) – this allows for an article before the adjective; a more specific search for examples that occur with the indefinite article could be achieved by replacing [pos="DT"] with [word="a"] – an adjective (JJ) and the word *impression*. Such a counter-intuitive method hinders corpus use and overwhelms many users.

Moreover, the content of many of the examples that the concordancer generates is too difficult for all but advanced students. Or many examples that are generated are not relevant: users might have to sift through hundreds of examples to find just one or two appropriate ones. For teaching and learning, tutors and their students need to be able refine their search by filtering concordance lines according to several parameters.



English Parts of Speech		
□ Noun (any)	□ Verb (any part)	☑ Adjective (general)
□ Noun (singular common)	□ Verb (base)	□ Adjective (comparative)
□ Noun (plural common)	□ Verb (past tense)	□ Adjective (superlative)
□ Noun (singular proper)	□ Verb (-ing)	
□ Noun (plural proper)	□ Verb (past participle)	
	□ Verb (-s)	
	□ Modal Auxiliary	
□ Adverb (general)	□ Pronoun (indefinite)	□ Wh- pronoun
□ Adverb (comparative)	□ Pronoun (personal)	□ Wh- possessive pronoun
□ Adverb (superlative)	□ Pronoun (possessive)	□ Wh- determiner
□ Adverbial particle		□ Wh- adverb
□ Conjunction (coordinating)	□ Determiner (general)	□ Possessive particle
□ Conjunction (subordinating)	□ Determiner (prefix)	□ Foreign word
□ Preposition	□ To (infinitve marker)	□ List item marker
□ Interjection	□ Cardinal number	□ Symbol
	□ Existential there	

Figure 1 *English Parts of Speech Window*



Query Editor

Word: make
○ this is a word
● this is a lemma
○ find the lemma of this word

Word: (Only POS tags set)
● this is a word
○ this is a lemma
○ find the lemma of this word

Word: (Only POS tags set)
● this is a word
○ this is a lemma
○ find the lemma of this word

Word: impression
● this is a word
○ this is a lemma
○ find the lemma of this word

Search description:

```
lemma "make" then
any word with part of speech tag "DT" then
any word with part of speech tag "JJ" then
word "impression"
within the same sentence
```

Query syntax: ● CQP ○ CSAR

Query string:

```
[lemma="make"] [pos="DT"] [pos="JJ"] [word="impression"] within s
```

Figure 2 *The Query Editor for multi-word searching*

Corpora must therefore be modified in two ways for LSP teaching. First, corpus tools need to be made simple so that non-specialists can use them. Second, more elaborate tagging is needed to show stylistic variation as well as other types of variation in language use. We have already taken steps to make our interface more user-friendly by replacing string codes with a new system whereby grammatical information is selected by ticking check-boxes (Figure 1) and we have simplified multi-word searching (Figure 2).[1]

3 Corpus-based approaches to vocabulary acquisition

The acquisition of a core discipline-specific vocabulary can provide a solid basis for delivering modules like business Russian, French for lawyers, Spanish for medics, and so on, whose lexicon is not covered by standard bilingual dictionaries or course books. Corpora can be used to generate frequency lists of the most common words and phrases in various LSP domains, and the benefits of this approach are clearly demonstrated by Butler (1974). Butler used frequency lists extracted from a small corpus (94,000 words) of systematically selected chemistry papers to "permit second-year chemistry undergraduates (or postgraduates), with no previous knowledge of German, to read papers from German chemical journals for comprehension and, where necessary, for translation" (50). Using exercises built around the most frequent words and collocations from his "chemistry" corpus, Butler aimed to equip his students with the requisite reading skills within ten teaching hours (plus 20 hours of independent study). The results were "gratifyingly successful" (53).

Quick acquisition of discipline-specific vocabulary is highly desirable in LSP teaching. Take, for example, the case of PhD students who need to learn a foreign language from scratch for their research. Intensive undergraduate ab-initio programmes are not suited to researchers' language needs; much of the material covered at Level 1 is of little relevance to researchers, while many important points of grammar and language use (from the researcher's perspective) are not covered at all. Gaining the necessary reading competence to understand academic texts is quite straightforward: academic texts are characterised by lexical and stylistic repetition; therefore, corpus-derived frequency lists are particularly effective. The same holds for grammar: teaching can be structured around the grammatical constructions that occur most frequently in academic texts. As specialised intensive language for research courses are not practicable at many institutions (especially in the case of less commonly taught languages) and there are concerns over the sustainability of established courses, new, cost-effective modes of delivery are highly sought after. A corpus-based, vocabulary-oriented approach is a

[1] http://corpus.leeds.ac.uk/it/

realistic and inexpensive way of enhancing reading skills quickly and effectively.

3.1 Frequency lists

Until recently, corpus users needed to ask computer specialists to compile frequency lists on their behalf. Nowadays, however, the compilation of frequency lists has become a standard feature on corpus interfaces, and users can use our tools to generate lists of single- and multi-word terms.

Single words	**Multiwords**
банк (bank)	ценные бумаги (securities)
предприятие (enterprise)	юридическое лицо (organisation)
кредит (credit)	денежные средства (monetary assets)
договор (contract)	федеральный закон (federal law)
товар (product)	заработный плата (salary)
рынок (market)	бухгалтерский учет (accounting)
финансовый (financial)	земельный участок (land plot)
налог (tax)	акционерное общество (public company)
страхование (insurance)	рынок ценных бумаг (capital market)
цена (price)	основные средства (fixed-capital assets)
учет (accounting)	фондовая биржа (stock exchange)
ценный (valuable)	процентная ставка (interest rate)
денежный (money-adj)	фондовый рынок (stock exchange)
имущество (property)	арбитражный суд (arbitrage)
налоговый (tax-adj)	недвижимое имущество (real estate)
прибыль (profit)	система управления (system of management)
государственный (state)	налоговые органы (tax authorities)
страховой (insurance-adj)	договор страхования (insurance contract)
стоимость (cost)	инвестиционный фонд (investment fund)

Table 1 *Russian keywords sorted by the Log-likelihood score of their significance.*

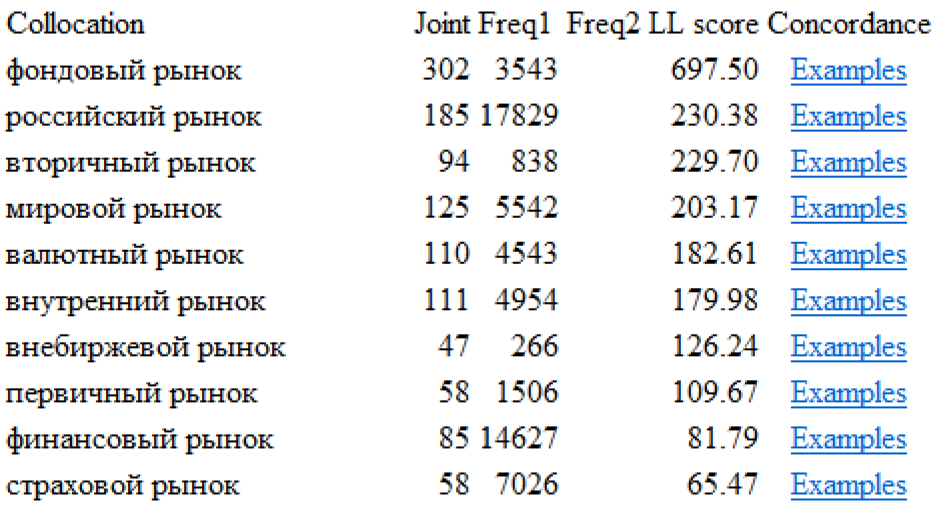
Single-word terms from specialised corpora are detected by Log-likelihood scores for their frequencies against a reference corpus (Rayson and Garside, 2000). An adaptation of the commonly-used Bootcat algorithm (Baroni and Bernardini, 2004) is used for the extraction of for multi-word terms. For example, we can start with the frequencies of Russian words in the overall Internet corpus (Sharoff, 2006) and the frequencies from the business corpus to get the list of keywords. Then the list can be extended by finding commonly used sequences of several words that contain at least one of the words in the keyword list. The importance of multiword units can be highlighted by non-compositional expressions, like *ценные бумаги* (securities) which literally means "valuable papers".

Besides using existing corpora to generate frequency lists, users may also compile and annotate (i.e. lemmatise and tag) their own collections of texts from, say, advertisements, CVs and covering letters, letters of complaint, political leaflets, etc. They can then extract the most frequent words and collocations from their corpora. This is a major benefit of corpus-based learning: while the lexicon of such a breadth of highly-specific domains cannot be represented in printed language-learning resources, which often take years to produce and generally cater for wider groups of users, corpus-derived frequency lists can be created within hours and, more importantly, can cater for the individual.

3.2 Automatic extraction of collocations

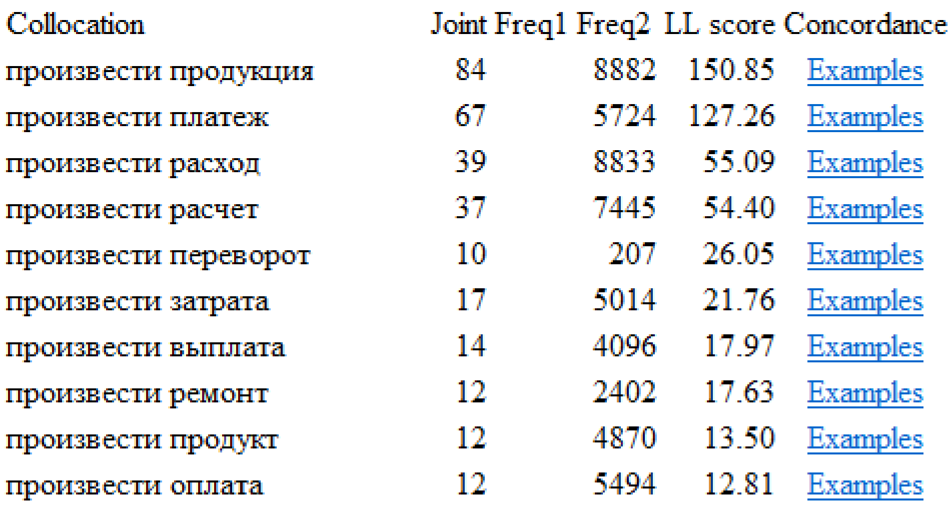
Tools for indentifying and displaying collocations in large corpora are now well established in many interfaces. There is an unresolved issue with selecting the most appropriate score to rank the collocates (Evert & Kren, 2001), but usually the Log-likelihood score is again a reasonable choice, as it takes into account the ratio of frequencies as well as the actual number of examples (Dunning 1993). Our tools can rank collocations either by their joint frequency or by several statistical scores (including Log likelihood).

Language learners often produce odd collocations, usually caused by interference from their native language (L1), and collocational information in most dictionaries is scant – obviously because space is limited. The automatic extraction of collocations is therefore another corpus-based function that can be used to good effect in learning and teaching and it exceeds the capabilities of existing printed resources. In keeping with the business theme of this paper, let us look at two examples from Russian: adjective + *рынок* "market" and *произвести* "to make, produce" + noun. Our tools allow users both to grammatically tag collocates and to select their position in relation to the search word; therefore, we can tell the corpus to show adjectives on the left of *рынок* (see Figure 3) and nouns on the right of *произвести* (see Figure 4). We used our Leeds-based Russian Business Corpus for both queries. To get a better idea of how these collocations are used in context students can then click on the "Examples" tab and view the concordances.



Collocation	Joint	Freq1	Freq2	LL score	Concordance
фондовый рынок	302	3543		697.50	Examples
российский рынок	185	17829		230.38	Examples
вторичный рынок	94	838		229.70	Examples
мировой рынок	125	5542		203.17	Examples
валютный рынок	110	4543		182.61	Examples
внутренний рынок	111	4954		179.98	Examples
внебиржевой рынок	47	266		126.24	Examples
первичный рынок	58	1506		109.67	Examples
финансовый рынок	85	14627		81.79	Examples
страховой рынок	58	7026		65.47	Examples

Figure 3 *Top ten adjective collocates to the left of рынок in the Russian Business Corpus*



Collocation	Joint	Freq1	Freq2	LL score	Concordance
произвести продукция	84		8882	150.85	Examples
произвести платеж	67		5724	127.26	Examples
произвести расход	39		8833	55.09	Examples
произвести расчет	37		7445	54.40	Examples
произвести переворот	10		207	26.05	Examples
произвести затрата	17		5014	21.76	Examples
произвести выплата	14		4096	17.97	Examples
произвести ремонт	12		2402	17.63	Examples
произвести продукт	12		4870	13.50	Examples
произвести оплата	12		5494	12.81	Examples

Figure 4 *Top ten noun collocates[2] to the right of произвести in the Russian Business Corpus*

[2] Collocates are displayed in their lemma form.

4 "Tagging for teaching"

Besides morphological and semantic tagging, now customary in many modern corpora, several other types of tagging can be used to enhance language learning. In this section, we look at genre classification and difficulty ranking.

4.1 Genre classification

Methods of tagging corpora are becoming more sophisticated and a major advance in corpus-based language teaching is the ability to annotate corpora to show stylistic variation. On the basis of previous research in which parameters for automatic genre detection and classification of texts from the Web (Sharoff 2007), we have been able to classify texts into such general genre categories as news items ('reporting'), legal texts ('regulations'), FAQs, advice ('instruction'), promotional materials ('advertising'), listings ('information') and everything else ('discussion').

For business Russian we have been able to apply formality tags so that business clichés can be extracted to help students differentiate between formal and neutral language use as well as to help them achieve a better understanding of register and to allow them to compile their own lists of generic phrases used in official writing. Such stylistic tagging has a much wider application and can be used to display other features such as non-standard or regional use.

4.2 Difficulty ranking

Although corpora have become a peripheral part of the teaching toolkit, many teachers have highlighted that their use is limited in that students find it hard to understand the content of the concordance lines. This is not surprising: corpora were not designed with the language learner in mind and texts were not selected according to how difficult they are to read. As a consequence, only advanced language learners have full access to corpus-based learning, and even they must sift through many examples to find just one or two appropriate ones.

Some recent advances have been made in this respect. Sharoff et al. (2008) established parameters for assessing the difficulty of texts and individual sentences in Russian (and other languages) and mapped the result to the CEFR levels. Methods have also been developed for automatically ranking concordance lines according to their difficulty by measuring lexical and syntactic complexity and sentence similarity (Segler, 2007; Kilgarriff et al., 2008; Kotani et al., 2008). These methods have been integrated into our system as additional options for sorting the concordance lines. This means that the difficulty level of the concordance lines displayed can be matched to the experience level of the students. Our user interface provides simple intuitive components, such as sliders, to ensure that setting the desired level of difficulty is easy. Even though, like other automatic processes, difficulty tagging is not completely accurate it still makes the tutors' task of selecting appropriate examples from the concordance lines much easier. Like all the processing features in our system difficulty tagging has been implemented in a plug-in fashion so that it can easily be replaced as better methods are developed.

5 Conclusions

We have shown in this paper several ways in which corpora can facilitate and enhance learning and teaching. A corpus-based approach to business Russian and to LSP teaching more generally has many benefits and helps to compensate for the lack of teaching materials, especially with respect to vocabulary acquisition and register recognition and differentiation.

Corpus-derived frequency lists can provide the basis for LSP courses and they can be used to meet the needs of *individual* language learners, unlike conventional course books that cannot cover vocabulary across the full breadth of LSP subjects. By uploading and annotating their own texts our users can generate frequency lists within their own area of research (even for very specific types of language such as that of instruction manuals or invoices). Corpora, not restricted by space limitations, can provide language learners with more comprehensive lists of collocations than printed dictionaries. A corpus-based approach to LSP teaching can also be used to facilitate register recognition and differentiation. Automatic genre classification allowed us to tag business clichés and formal expressions to help students extract set phrases and use them in their own sample official letters. Other tags can be used in other LSP domains to help learners differentiate between genres and registers more accurately.

Our study also shows that, although useful for LSP teaching, corpora require special tools and "plug-ins" to meet the needs of language learners. Such tools are being developed at Leeds and many other institutions, yet they are still in their infancy and are in need of further refinement. More research and collaboration between computer scientists and researchers in other fields is needed to ascertain the tools and functions that corpus users in various disciplines require. However, it is fair to say on the basis of existing evidence and expected technological advances that corpus-based tools will become easier to use and more functional, tagging will become more accurate and corpora will play an increasingly important role in research and teaching in various academic disciplines, including LSP teaching and language teaching more generally.

References

Baroni, M. & Bernardini, S. (2004). Bootcat: Bootstrapping corpora and terms from the web. In *Proc. of LREC2004*, Lisbon.

Butler, C. (1974). German for chemists. *Teaching Languages to Adults for Special Purposes (CILT Reports and Paper 11)*, 50-53.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1), 61-74.

Evert, S., & Krenn, B. (2001). Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. Toulouse, France.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proc. EURALEX'08*.

Kotani, K., Yoshimi, T., Kutsumi, T., Sata, I. & Isahara, H. (2008). EFL learner reading time model for evaluating reading proficiency. In *Proc. CICLING*, Haifa.

Segler, T. (2007). *Investigating the selection of example sentences for unknown target words in ICALL reading texts for L2 German*. Unpublished PhD thesis. University of Edinburgh.

Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. In *Proceedings of the Comparing Corpora Workshop at ACL 2000*. Hong Kong, 2000. Pp. 1-6.

Sharoff, S. (2006) Creating general-purpose corpora using automated search engine queries. In Marco Baroni and Silvia Bernardini, editors, *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna. http://wackybook.sslmit.unibo.it/

Sharoff, S. (2007). Classifying Web corpora into domain and genre using automatic feature identification. In *Proc. of Web as Corpus Workshop*, Louvain-la-Neuve.

Sharoff, S., Kurella, S. & Hartley, A. (2008). Seeking needles in the Web haystack: finding texts suitable for language learners. In *Proc. of Teaching and Learning Corpora Conference*, *TaLC 2008*, Lisbon.