Computer Assisted Individual Approach to Acquiring Foreign Vocabulary of Students Major

Nadezhda Almazova and Marina Kogan

Dep. of Linguistics and Cross-Cultural Communication,
St. Petersburg State Polytechnical University, St. Petersburg, Russia
almazovanadia1@ya.ru, m_kogan@inbox.ru

Abstract. Multiple challenges for organizing an effective ESP language course for non-linguistics post-graduate students at St. Petersburg State Polytechnical University (SPbSPU) are inherently rooted in the broad spectrum of students' majors in ESP classes. Diversity of students' academic interests calls for new approaches and for tailoring the course in accordance with the students' needs. Our study represents an approach to individualizing the course by introducing data-driven learning (DDL) elements into the syllabus. More specifically, our approach is aimed at having post-graduate students getting concordances of their readings corpora for identifying unfamiliar vocabulary. The paper describes the recommended software for concordance building, concordance-based activities with unfamiliar vocabulary and the way of controlling the vocabulary acquisition. Test results show steady progress in independent vocabulary acquisition among the experiment participants. Questionnaires show they see the usefulness and efficiency of DDL approach to identifying and learning unfamiliar vocabulary.

Keywords: data-driven learning (DDL) approach, teaching methodology, ESP course for post-graduates, concordance building software, knowledge rating, unfamiliar vocabulary.

1 Introduction

The problem of acquiring new lexis in specific purpose language (SPL) courses is traditionally in the focus of attention of both practical teachers, and methodology researchers [5, 17, 19, 24]. They argue the importance of forming teaching/learning strategies aimed at developing the students' ability to read words correctly, to know the meaning of a word within several different contexts, to use words both in reading and writing, as well as to use word-learning strategies.

Before answering all these questions a teacher, as well as students, should see quite distinctly what certain tasks they can fulfill without assistance but only mastering vocabulary. Basing on the research of Carver [3], Hu and Nation [15], and Chung and Nation [4] have come to the conclusion that at least 98 % coverage of the running words is needed for unassisted reading. Should it be acquired from extensive reading advocated by Cobb et al. [6], Day and Bamford [9], Hornst [11], and Pigada and

P. Zaphiris and A. Ioannou (Eds.): LCT 2014, Part II, LNCS 8524, pp. 248–257, 2014.

Schmitt [22], or is there any measurable learning from hands on concordancing [7] or an interactive on- line database, the idea supported by Horst et al. [12]?

The problem seems to be rather acute for Russian postgraduate students of Polytechnic University whose major covers different fields of sciences and humanities. Their study course comprises, as an obligatory part, taking an exam in the English language. The exam is mainly based on reading and comprehending special literature (scientific articles, monographs, etc.) which is demonstrated in adequate translation or interpreting. Adequacy implies the necessity of students' mastering their major vocabulary substantially in a pre-exam period at English classes. Translation is a common activity for many of them. A wide range of professionally -oriented lexis encourages intelligent uses of appropriate translation strategies as well as appropriate reading and test doing strategies. In this particular case a teacher must be very strategic about what vocabulary should be learned first, how a learning process should be organized, what word-learning strategies are preferable.

In spite of the fact that there have been several textbooks produced to teach academic vocabulary, examples are [16, 25] and papers devoted to technical vocabulary identifying and acquisition [4, 5], a teaching methodology here is of vital importance, because the problems a teacher faces are numerous. The situation turns out to be rather unpredictable when students with various major have English classes together. Not only their major is different, the level of English can be incomparable (from pre-intermediate to upper-intermediate). It is quite obvious that the learner-centered individual approach to learners in order to promote their interest to and the efficiency of learning their major vocabulary seems next to impossible to be implemented in a group of 20-30 students having classes once a week within a period of 90 minutes as it is often the case for ESP classes in post-graduate groups in Russian Universities [1]. All these factors predetermine some practical limitations to the current teaching context, and a vocabulary component of an ESP course seems to be quite challenging for a teacher, as far as all these issues must be taken into consideration.

One of possible ways of making the process of mastering major vocabulary more efficient is a tandem-learning described in one of our articles [23]. Another one is described here and is based on integrating computer in teaching context with the aim of eliminating those “postgraduate numerous class” minuses. It is based on DDL (data-driven learning) approach adapted to our circumstances. The main difference of our approach from what is described in the majority of publications on applying DDL in language teaching and learning is that we develop activities for concordances built by post-graduate students from their major reading corpora in English. We were inspired by T. Cobb and his colleagues’ idea of intensifying work with unfamiliar vocabulary through different activities with the concordance of the text which students have to read [6, 7, 13].

The activities depend on the features of the concordancer program available. For example, karTatekA allows analysis of a different lists (frequency list, inverted list, wordlengths list), lemmatization of words which provides an opportunity to unite all word family members in a single card and then to gain access to all contexts from the card, creating lexical and grammatical homonyms, word segmentation, and word

element and morpheme search etc. [1]. The serious obstacle towards its wide implementation is that it requires a very time-consuming procedure of preparing text in an original pdf format for building concordance with a sentence-length minimal context.

Being convinced that the idea of using concordances of reading corpora for intensifying vocabulary work is very promising and fruitful we decided to develop activities with different tools freely available on T. Cobb's *Compleat Lexical Tutor* website (http://www.lextutor.ca/) for consequent post graduate students' independent work.

Against this background the main research questions are thus:

1. Is the usage of this DDL adapted method efficient?
2. Does it reflect research interest or has it large uptake in practical teaching?
3. What is the learners' reaction to it?
4. Are the students largely successful in their outcomes?

2 Method

2.1 Features of Software Used

The concordance building software chosen was Text-based concordance tool from Comleat Lexical Tutor website developed and supported by Tomas Cobb at Université du Québec à Montréal (http://www.lextutor.ca/ http://www.lextutor.ca/ concordancers/text_concord/). It is a free-available on-line resource with a number of unique features considered by Diniz [10] and described in a series of T. Cobb and his colleagues' publications [e.g., 6, 12, 13]. Among the features not mentioned in the above papers but highlighted by Boulton [2], who found them typical of the most reputable websites, are its availability from any computer around the world via stable Internet –connection, its extremely low possibility of crashing, changing interface, moving site or being removed from the web. Even though Compleat Lexical Tutor does not offer the same conventional types of text searches as many other concordancer type programs do, and does not allow saving output concordance automatically on a hard disk, we decided to use this resource in our research because of

- simplicity of the original text (usually available in pdf format) preparation for getting its concordance;
- the maximum size of the text affordable for uploading to build concordance (up to 50 000 words) meets the needs of our post graduate students whose compulsory reading corpus of specialized papers consists of 47 000 words on average;
- unique features of the resource such as the *Text-based Range*, allowing users to upload up to 25 of their own texts and see how many of them each word appears in, and in which texts each word appears, and the *Vocabulary Profile* feature which analyses users' uploaded texts and compares their texts to the most-frequent-words-in-English word list and/or to the Academic Word List composed by Coxhead [8].

2.2 Analysis of Questionnaire Data

The data were collected from two achievement tests (Test 1 and Test 2) and the questionnaires completed by the experiment participants, post graduate students, totaled 6. The experiment participants were to build a concordance of their reading corpora using corresponding tools from *Compleat Lexical Tutor* website, to make up a list of 100 unknown words using a wordlist of the concordance, and to send it to the instructor. Basing on T. Cobb's and other researchers' works we could expect that the majority of unknown words will be of infrequent occurrence in the selected texts. To our surprise students reported from 30 to 50% words as unknown which are rather frequent in their corpora (with frequency more than 10). The recommended algorithm of studying the unknown lexis depended on the unknown word frequency in the student's corpus. For high-frequency words they were asked to read concordance lines and extended contexts from their papers to understand the meaning of the unknown word, and then to verify their guessing using a dictionary. For low frequency words they had to start with the same step, but taking into account a low probability of correct guessing based on one or two examples of the word usage, they were recommended to search for the word or word combinations using Multi-concordance tool on T.Cobb's website and in case of failure to find more examples there (which can be the case if the unknown item belongs to the specialist vocabulary) to use the *Google>books* query search before looking it up in a dictionary. As we showed [1] this is a very effective search tool for finding plentiful examples of usage of specialist lexis in domain-specific sources of usage in specialized and specialist contexts which are underrepresented in general on-line corpora such as BNC and COCA.

The questionnaires were completed after Test 2 because some questions implied the reflection on the Tests results. The respondents were from different departments, majoring in nanotechnology, physical electronics, physics of semiconductors, electric devices, economics, and finances; all but one of the respondents were male; the average age is 22. They all have been studying English for years, but have had different breaks in formal training. To decide if the usage of DDL adapted method was efficient we asked the students about their favourite strategies of memorizing new vocabulary. According to their answers, using an electronic bilingual dictionary and making up vocabulary lists were the most widespread strategies of dealing with unfamiliar vocabulary, "making no special efforts in hope that they will become familiar in a "natural" way (through reading, watching films, speaking to native speakers, etc.) sooner or later" was a second frequent choice; one third of respondents prefer using synonyms and guessing meaning from the context. The students use the same strategies reading different types of texts including specialized texts of their major. Usually most of them read specialized texts carefully just once, with only one student reporting that he does this twice. Nobody of the experiment participants has heard about DDL approach in language learning before the course.

The questionnaire asked students to rate strategies they used to clarify the meaning of the unknown words in their corpus according to the following five-point scale:

1 = never
2 = once or twice
3 = fairly often
4 = very often
5 = almost always

The results are presented in Fig. 1. The mean ratings on the ordinate axis show that for both high frequency words and low frequency words the students very often tried to guess their meaning from the nearest context (which means the concordance lines of their texts) or extended contexts. For low frequency words they also often used electronic mono- and bilingual dictionaries and *Google>books* query search, a new resource for them. Nobody used online corpora on a regular basis dealing with the unknown words from their texts. But many of them gave a try to this resource, new and unusual for them.

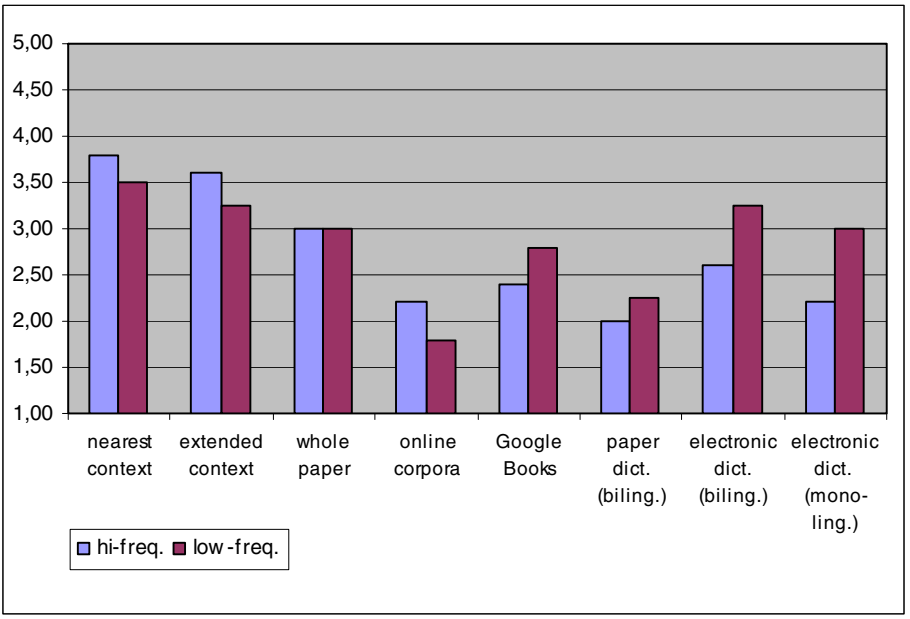

Fig. 1. Strategy use for learning unknown vocabulary for high- and low frequent words

They did pay attention to collocations of the unknown words with other words but often failed to notice grammar peculiarities (e.g., co-occurrence of articles, prepositions, commas, the place of words in the sentence, etc.). They reported that they did not encounter unknown words among off-list words produced by VocProfile tool. This means that the examples of the usage of single unknown words can be found in the BNC and COCA on-line corpora, and that the students know the specialist vocabulary well enough. However, the latter might be an illusive impression because the automotive word frequency range analysis deals with single words, not with word collocations. As a result, two- and multiword terms remain undetected.

Despite the multitude of examples of single frequent words usage (e.g., the number of tokens in COCA is 2048 for *turbine* and 175604 for *head*) there are no examples of the two-word terms from the specialist field (e.g. *turbine head* which is the difference between the static head and the losses through the installation) [1].

Some of other questions were in the form of statements on a 5-point Likert scale, from 1= strongly disagree to 5=strongly agree. They all found the concordance of their corpora which provides the wordlist helpful (M=4) in identifying unfamiliar vocabulary, the way of controlling and assessing the progress in mastering unknown vocabulary stimulating (3.6) and all but one would like to have similar tests during the whole course. They think that the activities with their texts based on using different tools from T. Cobb website are effective and useful for studying unknown vocabulary (4). There is only one "refuser" who did not like the method at all and is not likely to use it in future. Other participants are planning to continue to use this resource in their language studies after the end of the course. Taking into account the fact that this resource was absolutely new for them and training in the computer lab was very limited (actually only one academic session was allocated for the purpose) we feared that the students would find the resource difficult to use. To our surprise all participants except "the refuser" reported that it was not difficult to use the resource, with only one student pointing out that, overall, it is time-consuming. As for tools=activities most useful for learning unknown vocabulary their opinions divided. Text-based concordance and multi-concordance tools were the most frequent choice, with Text-based range, List_Learn and VocabProfile also mentioned. Two students are so enthusiastic about the DDL approach in language learning that they are ready to go further and master concordancer-type programs which can be installed on their PCs.

2.3 Testing Unfamiliar Vocabulary

Each of two tests contained 20 words selected randomly from 100 unknown word lists provided by the students. During the test they were to estimate a knowledge rating of the words according to the following scheme developed by Horst and Meara [14] and later used by Cobb and his colleagues [6].

0 = I don't know what this word means
1 = I am not sure what this word means
2 = I think I know what this word means
3 = I definitely know what this word means

and then provide translation equivalents for items they had given 3 and 2 points. Test 1 was conducted 2 weeks after they had identified 100 unknown words from their reading corpora. After another 3 weeks Test 2 was conducted. It contained items from Test1 which were given 0 or 1 points and those which were translated incorrectly. The rest words were selected randomly from the original 100 wordlist except for the words which had been included into Test 1. The students did not know the results of Test 1, and did not know which words we were planning to include into Test 2. So, they are supposed to have worked with all the words from their lists using different

strategies until Test 2. After completing Test 2 the students were asked to correct their translations from Test1 using the concordance or the context for the cases when they gave the correct translation of lemma but ignored a word morphology, e.g. *corrode* – *korrozia, which is a noun in Russian. By the next class they all submitted correct translations for these words. The results of the tests are presented in Tables 1 and 2.

Table 1. Word knowledge rating at Test 1 and Test 2

	Average number of words for each category from 6 lists of 20 words	
	Test 1	Test 2
0 (not known)	2.5	2.0
1 (rather unsure)	4.5	4.0
2 (less unsure)	4.5	2.6
3 (known)	8.1	11.4

Table 2. Translation results

	Average number of words translated correctly	Percentage of words with correct translation	Average number of words marked as "known"	Average number of "known" words translated correctly	Ratio of "known words" to words translated correctly from this category
Test 1	10.5	53	8.1	7.6	0.93
Test 2	14.4	72	11.4	10.2	0.90

The results show that the average number of words in the first three categories decreased while the number of words marked as "known" increased according to the students' self- evaluation. The test results prove that their self-evaluation is correct: they translated correctly 91.5% words marked as "known". Back to the questionnaires, they think the results of the tests are valid (3.8) for making conclusions about the degree of learning the rest of the words from the given group (of 100 words). They estimate that they do not know about 0.5-2% words in their reading corpus, which is within 230 – 950 words for their corpora of an average size of around 47 000 words.

We did not take into account the results of the "refuser" who did not use DDL approaches while working with the unknown vocabulary. We have good reasons to suspect that he did not use his favourite strategy (bilingual dictionary) either, because none of the words from Test 2 was marked as "known" by him, and only 3 words were translated correctly. On the other hand, he marked as "not known" only 3 words. This can be regarded as an illustration of the conclusion – very encouraging, in our opinion – made by some researchers that there is the learning impact of even one or two encounters with a new word (though to capture it, very sensitive measures are

required) [13]. We used so called "active recall" [18], the most difficult for learners way of testing and, probably, the only possible for us having to deal with lists of 100 words each. Our "refuser" definitely encountered unfamiliar words at least a couple of times, selecting them from the wordlist and then copying them into a special file.

2.4 The Analysis of Unknown Vocabulary

We have conducted the analysis of the vocabulary that our students are unfamiliar with using the software from VocabProfile section of *Compleat Lexical Tutor* website. The results are presented in Fig. 2. Curve 1 shows the distribution of all words from the students' 100 word lists of unknown words, a total of 611 tokens. Curve 2 shows the distribution of words found in more than one of 100 word lists. The total number of such words is 70, including 9 words unfamiliar to the half of the students.

According to the plot most of the first hundred unknown words from the students' reading corpora belong to the first – fifth thousand frequency bands of the *most-frequent-words-in-English* BNC-COCA word list. The majority – 185 tokens or 153 words excluding re-occurrences for curve 1, and 32 tokens for curve 2 – belong to the third band. The possible recommendations could be as follows.

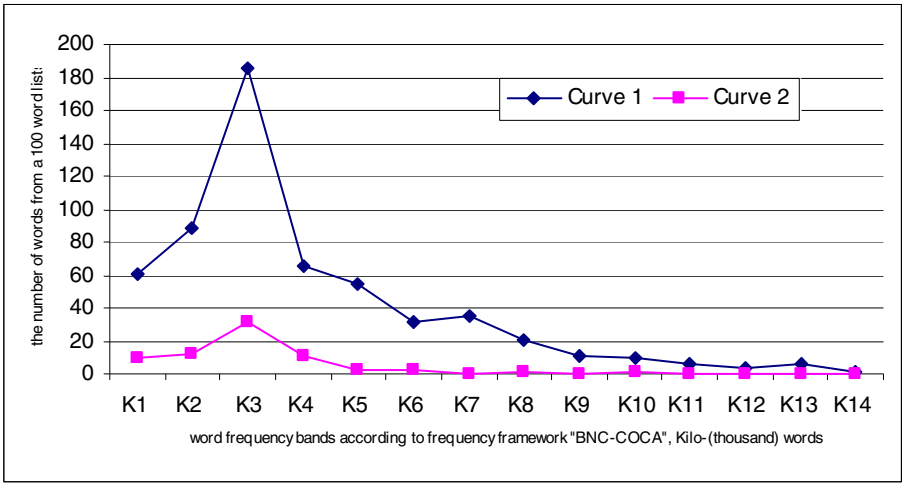

Fig. 2. The distribution of the experiment participants' unknown words according to the frequency framework «BNC-COCA»

The students could be recommended to revise the first – forth thousand frequency band word lists from List_Learn section of the T. Cobb's website. It allows users to detect unknown words looking sequentially through the word list of a given frequency band. Then the meaning of an unknown word could be understood from a number of BNC examples, with the most telling of them being stored on the user's PC. Hopefully by the end of the course post-graduate students will be able to expand their

vocabulary till 5 000 General English words and collocations in accordance with the Russian National Standards requirements, following this algorithm independently. The words unknown to different students are to be accumulated in a special bank so that foreign language teachers could develop vocabulary training exercises and use them in the ESP post-graduate classroom.

3 Conclusions

The small sample in the present study needs extending though the results so far have been promising in terms of what learners do with their reading corpora to identify the unknown vocabulary and learn new words using DDL approach which is new for them. They are largely successful in their outcomes and the progress in acquiring new vocabulary proven by tests. It is important to notice that they achieved these results focusing on the unknown words, not re-reading carefully their texts several times as participants of a similar experiment described in [6] did.

The described algorithm could be recommended for introducing this DDL adapted method into the ESP post-graduates course for organizing their individual work in the course taking into account generally positive feedback, small amount of special training required, and a relative simplicity of the control procedure. A more detailed analysis of students' vocabulary needs based on the unknown vocabulary from their reading corpora is required.

References

1. Almazova, N., Kogan, M.: Organizing polytechnic post-graduate students individual work on required reading corpora (wihtin ESP course). Universitetskii Nauchnyi Zhurnal=Humanities and Science University Journal 6, 13–25 (2013)
2. Boulton, A.: Beyond concordancing: Multiple affordances of corpora in university language degrees. Procedia – Social and Behavioral Sciences 34, 33–38 (2012)
3. Carver, R.P.: Percentage of unknown vocabulary words in text as a function of the relative difficulty of the text: Implications for instruction. J. of Reading Behavior 26(4), 413–437 (1994)
4. Chung, T.M., Nation, P.: Technical vocabulary in specialized texts. Reading in a Foreign Language 15(2), 103–116 (2003)
5. Chung, T.M., Nation, P.: Identifying technical vocabulary. System 32(2), 251–263 (2004)
6. Cobb, T., Greaves, C., Horst, M.: Can the rate of lexical acquisition from reading be increased? An experiment in reading French with a suite of on-line resources. In: Raymond, P., Cornaire, C. (eds.) Regards sur la didactique des langues secondes, Éditions logique, Montréal, pp. 133–153 (2001)
7. Cobb, T.: Is there any measurable learning from hands-on concordancing. System 25(3), 301–315 (1997)
8. Coxhead, A.: A new academic word list. TESOL Quarterly 34(2), 213–238 (2000)
9. Day, R.R., Bamford, J.: Extensive Reading in the Second Language Classroom. Cambridge University Press, Cambridge (1998)

10. Diniz, L.: Comparative review: Textstat 2.5, AntConc 3.0, and Compleat Lexical Tutor 4.0 Language Learning & Technology 9(3), 22–27 (2005)
11. Horst, M.: Learning L2 vocabulary through extensive reading: A measurement study. Canadian Modern Language Review 61(3), 355–382 (2005)
12. Horst, M., Cobb, T., Nicolae, I.: Expanding academic vocabulary with an interactive on-line database. Language Learning & Technology 9(2), 90–110 (2005)
13. Horst, M., Cobb, T.: Growing academic vocabulary with a collaborative online database. In: IT-MELT 2001: Information Technology & Multimedia in English Language Teaching, Kowloon, Hong Kong (2001)
14. Horst, M., Meara, P.: Test of a model for predicting second language lexical growth through reading. Canadian Modern Language Review 56(2), 308–328 (1999)
15. Hu, M., Nation, I.S.P.: Vocabulary density and reading comprehension. Reading in a Foreign Language 13(1), 404–430 (2000)
16. Huntley, H.: Essential Academic Vocabulary. Houghton Mifflin, Boston (2006)
17. Joe, A., Nation, P., Newton, J.: Vocabulary learning and speaking activities. English Teaching Forum 43(1), 2–7 (1996)
18. Laufer, B., Goldstein, Z.: Testing vocabulary knowledge: Size, strength, and computer adaptiveness. Language Learning 54(3), 399–436 (2004)
19. Nation, I.S.P.: Learning Vocabulary in Another Language. Cambridge University Press, Cambridge (2001)
20. Nation, I.S.P.: Teaching Vocabulary: Strategies and Techniques. Thompson Heinle, N.-Y. (2008)
21. Nation, P., Chung, T.: Teaching and testing vocabulary. In: Long, M.H., Doughty, C.J. (eds.) The Handbook of Language Teaching, pp. 543–559. Blackwell Publishing Ltd., Hong Kong (2009)
22. Pigada, M., Schmitt, N.: Vocabulary acquisition from extensive reading: a case study. Reading in a Foreign Language 18(1), 1–28 (2006)
23. Popova, N., Kogan, M.: Didactic Links as means of Profiling Technology Realization for Masters of Linguistics. In: Proceedings in the 1st Global Virtual Conference workshop, pp. 185–189. Publishing Institution of University of Zilina, Zilina (2013)
24. Read, J.: Assessing Vocabulary. Cambridge University Press, Cambridge (2000)
25. Schmitt, D., Schmitt, N.: Focus on Vocabulary. Longman Pearson Education, White Plains, NY (2005)