Using Language Corpus in Teaching Foreign Language Vocabulary

Ewa Donesch-Jezo

Jagiellonian University,
Krakow, Poland

Paper Received on: 15/01/2013
Paper Reviewed on: 20/01/2013
Paper Accepted on: 21/01/2013

Abstract

Teaching of L2 vocabulary has regained interest in applied linguistics in recent years. It is mainly due to the development of new ‘word-centered’ approaches to L2 language teaching, as well as to the access to the internet databases of words (language corpora) which comprise a limitless number of various types of vocabulary used in authentic texts. Language corpora can also be used as invaluable sources of various activities for teaching L2 grammar and vocabulary. What can teachers do to facilitate their learners’ acquisition of new words?

This paper aims to give the answer to this question by providing a number of examples of learning activities showing how the teaching of vocabulary items can be realized in the L2 classroom in an effective and attractive way with the use of language corpus together with concordancing software. The first part of this paper presents some theoretical information on processing, storing and producing new words, whereas the second part offers learning tasks which aid at vocabulary memorizing. The last section of this paper shows how to teach learners the ways to cope with gaps in their vocabulary.

Keywords: vocabulary acquisition, language corpus, concordancing, vocabulary learning tasks

Introduction

The revival of the interest in vocabulary teaching in the last 20 years has been ascribed to two factors: 1) accessibility of computerized databases of words (language corpora), and development of word centred or lexical approach in vocabulary instruction.

For many years learning lexis was secondary to grammar learning in language acquisition due to the assumption held by some researchers (Larsen-Freeman, 2000; Zimmerman, 1997) that once students learnt grammatical structures, lexical items to fill the slots in syntactic frames could be learnt later according to the students’ needs. In the era of autolingualism, which emphasized learning grammatical and phonological structures, vocabulary learning was left to take care of itself (see Schmidt, 2001).
Recently this view has changed dramatically, and more and more researchers (e.g. Coady and Huckin, 1997; Thornbury, 2002) started to see that lexical competence is central to gaining the ability of effective communication. In order to communicate successfully and appropriately a certain number of words must be acquainted. As Wilkins (1976) stated, “Without grammar very little can be conveyed, without vocabulary nothing can be conveyed.”

The questions which arise in relation to vocabulary acquisition are: What does it mean to know a word? How are words learnt? How many words and which ones should one know to communicate effectively? What can teachers do to facilitate their learners’ acquisition of new vocabulary?

Learning the L2 vocabulary entails the following challenges for the learner:

- making the correct connections between the form and the meaning of the words;
- distinguish the meanings of phonologically closely related words, e.g. noon and moon;
- when producing language, to use the correct form of a word for the meaning intended, e.g. law, not low;

To conform to these challenges, the learner needs to:

- acquire a great number of words for use in understanding and producing language;
- remember words, and be able to recall them easily;
- develop strategies for coping with gaps in word knowledge, including coping with unknown words.

This paper presents both a theoretical and practical basis for the methodology which underpins the teaching and learning of vocabulary and considers some of the methodological implications for language teachers.

1. Vocabulary learning and teaching

The learning of a second language (L2) vocabulary is obviously different from the learning of the first language (L1) vocabulary. Second language learners already possess the knowledge of the words of their first language, as well as the system of concepts that these words are related to. They also possess the network of associations that link the words to one another, for example, words like red and white, or hand and leg are linked together as the members of the same category of concepts. The learning of an L2 vocabulary involves both
learning a new conceptual system, and constructing a new vocabulary network – a second mental lexicon (Saville-Troike, 2006). The acquisition of new words may proceed explicitly (consciously) with the aid of the teacher’s explanations or implicitly (unconsciously), which is facilitated by exposure to language input, for example, extensive reading of various texts. The role of the teacher in implicit learning is to provide the students with suitable input and to equip them with the strategies of inferring the meaning of words from their context.

Another difference between L1 and L2 vocabulary learning lies in the number of words acquired by native speakers and by second language learners. An educated native speaker knows vocabulary of about 20,000 words or word families, whereas most adult second language learners during several years of L2 learning usually acquire about 5,000 words or word families (DeCarrico, 2001). The number of words the learner should know depends on his or her needs. Apparently, different number of vocabulary is required for those who plan to go on a holiday to an English-speaking country, and for those who intend to study in a British university. Researchers, such as Nation (1990), Meara (1995) and Carter (1998), argue that a threshold level or a “core vocabulary” of 2,000 high-frequency words should be taught using explicit instruction and appropriate exercises as soon as possible after the beginning to learn an L2. This number of words is approximately what native speakers use or hear in daily situations. Having acquired 2,000 words, the learners can acquire most of low-frequency words implicitly during listening or reading of various texts, and also they can more easily infer the meaning of words from the context. Learners who intend to study in an English-speaking country need to acquire another thousand of high-frequency words plus the strategies of coping with low-frequency words which they encounter when studying their mainstream subjects.

Learning words means accumulating them in the learner’s memory and remembering them. Since frequent encountering of a given word in various texts is crucial in both explicit and implicit learning, students should be exposed to an ample of authentic texts. This enables them to meet vocabulary in a real context, which facilitates its consolidation and memorization. In the teaching/learning of a new word, a significant role is attributed to learners’ noticing it in the text (Schmidt, 1995). It initiates its mental processing, which facilitates the word’s transfer from short-term memory to long-term memory. Short-term memory stores only a limited number of vocabulary items for a short time (1-2 seconds), whereas long-term memory stores unlimited number of words for an unlimited time. For
words to be integrated into long-term memory, they need to undergo different kinds of
cognitive operations in working memory, such as reasoning, understanding and learning.

The research conducted by Schmidt (1990, 2000) showed that noticing a new word is
crucial for its perception, and that it is related to the frequency with which a new word
appears in the input. The activities enhancing a word transfer from short-term memory to
long-term memory involve its consolidation by repetition, elaboration or manipulation. The
more the student manipulates with the word, while performing different kinds of exercises
and tasks, the more effective its incorporation to long-term memory is.

Schmidt (2000) emphasizes the role of techniques which by directing of the learners’
attention enhance the perception of new vocabulary items. They include marking new words
with colour markers, using fonts of various size and underlining new words. Marking new
vocabulary in the text is beneficial for its learning, since it helps not only to notice it, but also
to observe its form and relations with other words.

2. Explicit and implicit vocabulary learning

New words should be presented in context, that is, in connection with the situation in
which discourse is produced. In explicit vocabulary learning, students are engaged in
activities that focus their attention on words and their connection with other words in the text
(DeCarrico, 2001, p. 286). As Nattinger (1988, p. 69) states, “the meaning of a word has a
great deal to do with the words with which it commonly associates.” The knowledge of
vocabulary, therefore, means more than the knowledge of the meaning of a given word in
isolation; it also involves the knowledge of the words that co-occur (collocate) with it.
Collocations are pairs or groups of words that co-occur with very high frequency, e.g. *rancid
butter*, *sour milk*, *sleep soundly*. The lack of the knowledge of collocations results with
producing deviant or odd speech or writing, such as *“spoiled butter”* or *“sleep healthily”*.
Collocations may be grammatical or lexical. In grammatical collocations some verbs or
nouns co-occur with particular prepositions, e.g. *depend on*, *abstain from*, *by accident*, *in
retrospect*, or some verbs co-occur with a particular complement, e.g. *make an appointment*,
*and take an examination*. In lexical collocations there is a combination of words having a
particular semantic content e.g. *catch a cold*, *effective therapy*, *laugh loudly*, *deeply
absorbed*. On the one hand, collocations, because of their unpredictability, may cause
problems with their learning, but on the other hand, as Nattinger (1988) says, collocational
association’s aid students in the positioning of these words in memory and in creating the
semantic area of a word. Therefore, using by teachers texts containing collocations and explicit directing the student’s attention to them and explaining the semantic associations between the words being the collocation components, as well as presenting new words in the most common collocations they form, seem to be an effective method of vocabulary acquisition. Classroom activities designed for this purpose include word-matching, in which students from a list of words (e.g. adjectives) choose those which collocate with a given word (e.g. a noun), or gap-filling exercises, in which students are asked to choose all possible words from a provided thematically related list. However, according to some researchers, e.g. DeCarrico (2001) and Thornbury, 2002), collocations should not be included in an early stage of language learning. In this stage, learning should focus on developing a large number of basic vocabulary in typical contexts in which the words are usually found.

The use of language corpora, which are nowadays accessible online, allows students to check the probability of the co-occurrence of two words. For teachers language corpora are a useful source of linguistic data for preparing teaching activities. Data obtained from a language corpus allow determining the criteria of word associations in collocations. For example Stubbs (1995), analyzing a language corpus, noticed that the verb “cause” collocates with nouns meaning something unpleasant, such as problem, trouble, damage, and the verb “provide” collocates with words meaning something positive, such as aid, assistance, support.

Another common multiword units which allow learning vocabulary in frequently occurring combinations with other words are lexical phrases. Nattinger and DeCarrico (1992) define lexical phrases as “chunks” of language of varying length and conventionalized form which frequently occur in our everyday life. Some of them have fixed composition, e.g. By the way, how do you do? Give me a break, and some have a basic frame with slots to be filled, e.g. a ... ago, the ...er X, the ...er Y. Lexical phrases are types of collocations and are more or less idiomatic. Yet they differ from idioms in that they fulfill a particular discourse function, such as expressing time, greetings, relationship between ideas, e.g. how are you?, see you later, would you mind ...? In written discourse achieving cohesion without using lexical phrases is impossible. Some phrases have the function of discourse markers signalling, for instance, contrast (on the other hand; but look at), addition (in addition to) or exemplification (to give you an example; in other words). Teaching lexical phrases is advantageous for learning various types of discourse, since they are stored and retrieved as whole chunks, allowing learners for the use of phrases that they may not be able to create by
themselves properly. In teaching lexical phrases a crucial role has their presentation in discourse and the analysis of their functions in a given text.

One of important considerations, related to considerable maximization of vocabulary, is teaching word families. A word family is a set of words that includes a base word (e.g. work) plus its inflections and/or derivations (e.g. work+s, work+ed, work+er, work+ing). Research shows that in the mind these different forms of the same word are grouped together.

Thus instead of teaching a single word form, teachers can use exercises consisting of creation or derivation of word families out of a given word.

Implicit vocabulary learning which is mainly achieved through exposition to various types of texts is significantly facilitated by the use of the strategy of guessing word meaning from context. This strategy is extremely important in acquisition of low-frequency vocabulary, such as specialist words in authentic texts. It is indisputable that possessing background knowledge about the topic of discourse aids inferring the meaning of unknown words, as it provides a framework for their incorporation with information already known. Even if the learners do not have such knowledge, the teacher can equip them with the strategies of guessing word meaning. One of the strategies is that proposed by Clarke and Nation (1980), based on learners’ examination of the immediate neighbourhood of an unknown word. The linguistic clues signalling the meaning of the unknown word can be its part of speech (e.g. noun, verb, adjective) and its immediate context, that is, the words preceding and following it, such as connectors (but, because of, when) and adverbs (however, as a result), which indicate coordinating or subordinating relation of words in compound sentences. Another helpful clues are types of rhetorical relationship, such as cause and effect, contrasting, inclusion, exemplification and summarizing, as well as punctuation and presence of reference words, such as this, that, it which provide useful information.

After the presentation of a new lexical item in a written or spoken text, learners need to practice recognizing, manipulating and using it in produced by them texts. The practice is based on the idea of using the word in activities resembling real situations, such as:

- discussions, problem-solving, role-play, which require the use of new lexical items;
- using new lexical items in written tasks.

A variety of exercises and tasks are proposed by authors, such as Scrivener (1994), Carter (1998), DeCarrico (2001), Nation (2001), Thornbury (2002). They include:
• matching pictures to lexical items;
• matching parts of lexical items to other parts, e.g. beginnings and endings;
• matching lexical items to others, e.g. collocations, synonyms, antonyms, lexical phrases;
• word-building – using prefixes and suffixes to build new lexical items from given words;
• classifying items into word families;
• using given lexical items to perform a specific task;
• filling in crosswords, grids or diagrams;
• filling in gaps in sentences;
• memory games.

3. Language corpus in vocabulary teaching

A useful tool in teaching vocabulary is analysis of corpus data. It provides valuable information for both students and teachers about how language is used in real-life situations. A corpus is a collection of authentic texts (written or spoken transcripts) that are stored in an electronic form (Partridge, 2006: 1). Its size can range from a few sentences to millions of words. Linguistic information is typically presented in the form of concordances (Tribble and Jones, 1997). A concordance is a list of all the occurrences of a particular word or phrase in a corpus, presented within the context (usually a few words to the left and right of this word). Concordances are obtained using the software called a concordancer.

One of the first teachers who used a concordancer was Tim Johns, who was the author of the Data Driven Learning (DDL) (Johns, 1991). DDL is an approach to language learning based on the assumption that the use of authentic language together with a concordancer will enable the learners to observe the language as it is used in real-life situations. What is more, in DDL the learning process is based on the learner’s discovery of rules and patterns of language use.

3.1. Study design

The study included 30 students in the fourth year of medical studies – the participants of two classes of their obligatory ESP (English for Specific Purposes) course at the Medical College of Jagiellonian University in Krakow. The students’ level of proficiency in English
was intermediate and upper intermediate. The duration of an experiment was 6 hrs, scheduled 2 hrs weekly.

The corpus included 10 research article Introductions selected from four reputable Anglo-American journals in the field of medicine. All these texts had been retrieved online from the Jagiellonian University library. The selected articles were written by native speakers of English. The students, organized in pairs performed the analysis of the texts to find the localization and the use of some key lexical items. The analysis was conducted manually by reading the texts and recording the occurrence of particular lexical items in the Introduction section of research papers and with the use of concordancing software *MonoConc Pro 2.2*.

3.2. Method

Before the analysis of the corpus, the teacher explained to the students what a concordance is, how the corpus is made and showed them how the concordancer works. The students were also explained what the function of the Introduction of the research paper is. Then the students, working in pairs, received a corpus of 10 Introduction sections of research papers along with a set of language learning exercises designed by the teacher (the author of this paper). The aim of these tasks was to guide the students’ analysis, and to make them aware about the use and function of some important vocabulary items present in the text. These tasks were discovery activities, which drew the students’ attention to the functions and form, as well as associations formed by these words. The tasks included open-ended questions, gap-filling, matching the words and filling the table with words that form collocations with the given key word. Some examples of open-ended questions is given below.

- What words do the writers use to present their research?
- Why do the writers in some sentences use “we”-subject (what do they say in these sentences)?
- Which words are most frequently used by the writers to tone down the assertiveness of their statements (check the concordances for the words, such as *can, could, may, might, seem, possible, likely*)?
- Where are the words *however, therefore, moreover* located in sentences?
- Which words are most frequently used by the writers to present the aims of their research?
The students worked either on the data shown on a computer screen, or they worked on the printouts of concordances produced during class. The teacher supervised the students’ work and provided a feedback. The examples of tasks performed by the students are presented below.

In order to help students organize the information which they obtained from the concordance program, the students wrote their observations and formulated their conclusions.

Table 1. Students’ worksheet

<table>
<thead>
<tr>
<th>KEYWORD</th>
<th>Part of speech:</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEFT SORT</td>
<td>Write the words (or word classes) which appear frequently before a keyword.</td>
</tr>
<tr>
<td>RIGHT SORT</td>
<td>Write words (or word classes) which appear frequently after a keyword.</td>
</tr>
</tbody>
</table>

Can you formulate any conclusion? Write it below.

3.3. Corpus-based exercises

Before creating concordance lines, the students created the wordlist which showed the frequency of occurrence of all the words occurring in the text. A fragment of such frequency list is presented in Figure 1.
By creating a frequency list before analyzing concordances for certain key words, it is possible to see which lexical items are most common in the texts in a particular domain. These words can be chosen for further study. The obtained frequency list (Figure 1) shows that after the functional words, such as prepositions and articles, among the words of high frequency in the studied corpus are nouns, such as patients, risk, disease and the pronoun we. They can be the focus of further study.

In order to find the collocations which the noun “patients” form in the Introduction section, the students were asked to produce a concordance on the key word “patients” and to write in the worksheet the words occurring on the right of this key word. The concordance lines obtained is shown in Figure 2.
Figure 2. Screenshot from MonoConc Pro showing the concordance for the keyword “patients”

From the key word displayed in this way, the students could list the words that collocate with “patients”, notice that prepositions, such as with, in, on are used in these combinations, and see which adjectives precede this word.

In order to check which word is more often used as the subject of sentences describing research procedures: the first person pronoun “we” in active-voice sentences, or the noun “study” in passive-voice sentences, the students created concordances for the key words “we” and “study”. Figure 3 provides the concordance for the keyword “we”.
On the basis of obtained concordances, the students could observe and compare which of these two words occur more often: whether the number of matches for active-voice sentences with the “we”-subject is higher than the number of matches for passive sentences with the subject “study”, or the other way round. The students could also notice which words these words collocate with, which allowed them to see what kind of the writers’ research activities are described in the passive voice and what - in the active voice. After making notes, students could pass on to make generalizations and conclusions, where possible.

In order to obtain this kind of information, students do not need a computer and a concordancer. They can make their own concordance lines by writing out all the sentences or parts of sentences that include the examined words, however, making concordance lines manually is not as quick as with the use of a concordancer. Another possibility, which the use of a corpus and a concordances gives to teachers is editing (printing) concordances. This gives teachers the possibility of designing exercises consisting of the concordances in which the key word was blanked out, and the students’ task is to guess what the missing word is.
4. Results and conclusions

The above section has presented the possibilities of the corpus-based teaching/learning in vocabulary acquisition. The examples of exercises designed for students showed the nature of a corpus-based approach, in which the concordances seen on the computer screen or on their edited version allow the students to examine the vocabulary in their natural context, and with the teacher’s feedback to draw generalizations. The effectiveness of this approach was confirmed by high scores the students obtained in their vocabulary tests after finishing the vocabulary-learning session as compared to the scores obtained before it.

Informal conversations with the students showed that they considered this type of vocabulary learning as ‘an interesting experience’, and that this knowledge helped them in their understanding of scientific texts.

To conclude, a corpus-based approach to the raising of vocabulary awareness, combined with the use of a concordancer and learning tasks, which guide the students’ vocabulary examination can, it may be argued, be effectively used in foreign language classes at university level. The main advantages of this methodology are:

1. It promotes discovery-based learning by encouraging the students to discover linguistic ‘facts’ by themselves, which is in accordance with the principles of Data Driven Learning.
2. It uses authentic texts (including scientific texts), which are particularly rich in vocabulary learning potential.
3. It provides the students with the vocabulary, the knowledge of which is useful in understanding of the scientific texts they are expected to read.
4. Examining a specialist corpus helps the students to understand that a particular text type has its own characteristic vocabulary.
5. The print-outs of the concordances can be used in the preparation of teaching materials and tests.
6. Language corpora are valuable vocabulary input, providing information on the frequency and collocations of the words in authentic texts.
7. With the use of a corpus and concordancing programmes the discovery of this information is possible not only by the lexicographers but also for students in the classroom.
References


