A Software System for Learning the Vocabulary and Collocations: Results of a Training Experiment

Pavel Diachenko

Institute for Information Transmission Problems (Kharkevich Institute), RAS
Bolshoi Karetny 19, Moscow, 127994 Russia
pavelvd@iitp.ru

Abstract

This paper[1] describes the progress of our work presented at m-ICTE2006 (Boguslavsky et al., 2006; Diachenko, 2006). The general purpose of the work is to develop a software system which promotes the mastery of the combinatorial potential of natural language words by language learners. It lists some basic concepts of *Lexical function* theory and describes the architecture and functionalities of the software system developed. We describe the idea and the results of training experiment.

1. General concept

Good command of idioms and collocations is an important ability of any natural language speaker. This ability needs to be developed not only in foreign language learners but also in people wishing to enhance their linguistic competence in the native language. Since systematic description of idiomaticity is still a very difficult issue and linguistic resources focused on idioms and collocations are scarce, tools intended for boosting up this ability are poorly represented in modern language learning techniques. We developed a software system which can be used to improve human knowledge of the combinatorial potential of words.

The idea of a software system which should facilitate the mastering of LF was first put forward by Juri Apresjan in early 1990s, who compiled a prototype dictionary for English and Russian underlying such a system, to be followed in a mid-1990s INTAS-funded project, CALLEX (Computer-Aided Learning of Lexical Functions), where German was also incorporated. Both initiatives used a combination of linguistic knowledge and innovative linguistic technologies for teaching the idiomatic aspect of the lexicon. It was largely based on LFs and the theory of *lexical decomposition* proposed by Ju. D. Apresjan (Apresjan et al., 2002; Apresjan, Tsinman, 2002).

Since then certain aspects of LF theory have progressed, and the material of Russian and English dictionaries has been drastically updated. We took over all of the general ideas of the CALLEX project; also, we use the newest version of the dictionary developed by its authors as basis for our work.

A closer look at the basic linguistic concepts mentioned above and lexical games was taken in (Boguslavsky et al., 2006; Diachenko, 2006). Here we will give a short summary.

2. Lexical functions

An elementary *lexical function* (LF) is a relation between one word or word combination, which is called *function argument*, and another word or word combination, which is called *function value* corresponding to this argument. Not every lexical function value can be predicted with a 100% accuracy by its argument because it is not completely motivated semantically. In a general case all lexical functions have multiple values.

Example: **Magn -** *a large degree or a high intensity of X.*
Magn (CONTRAST (NOUN)) = *sharp / harsh / startling / striking*,
Magn (IMAGINATION) = *lively / warm / heated*,
Magn (SILENCE) = *deep / profound / complete / dead / deadly / mortal.*

By now the dictionary of the system contains a description of nearly 120 LFs. According to I. A. Mel’čuk’s hypothesis, there are nearly 55 elementary lexical functions, which may additionally form *compound lexical functions.* Most of these functions are universal for all languages. Both elementary LFs and compound LFs are called *standard.* Also, there is a large number of non-standard LFs which are not universal (not language-independent). In the system we are presenting, only standard LFs are used.

According to the logical division of the dictionary, we can separate one game operating with the material of the LF dictionary from those operating with the material of the word dictionary. These dictionaries are currently independent of each other because we find it important to separate two different sets of LF values - the set of values which represent its main features and values which just correspond to the entries in the word dictionary. The second set of LFs, among many non-trivial values, contains also trivial and largely repeating values. For example, most *fruits* and *vegetables* have a value for **Bon** (a standard positive evaluation of X) *mellow*, **Ver** (the property which is normally required or expected of X) *ripe*, **Func0** (X exists or is taking place) *grow(s)* and **Real1** (To use X according to its destination) *eat.*

3. Lexical games

The purpose of our work is to develop modern software that can be used for lexicon learning. The process of

[1] This work was partially supported by the Russian Foundation of Fundamental Research (grant No. 08-06-00344).

learning is organized in the form of linguistic games, the current number of games is 5. In our training experiment we used only the LF dictionary.

An elementary step of all the games is one question. For each answer given by the user, the system estimates the number of points won. The more difficult the question, the more points are given for a correct answer. Some of the questions allow for more than one correct answer. Additional answers for such questions can bring extra points to the user. The system determines the baseline of the user's knowledge as his or her ability to provide at least one correct answer to every question offered. Comparing the number of points won with the baseline the user can measure his or her study progress.

All new given answers which were not presented in the dictionary are stored by the system. They can be retrieved by the teacher for analysis of the user's study progress as well as by dictionary authors to improve the dictionary.

We believe that the developed system can be useful both in improving the level of mother tongue and for studying the idiomaticity of a foreign language. It can be improved by adding new linguistic games, learning techniques and the development and usage of new support utilities.

4. The training experiment

We made an experiment to measure the study progress of people working with our system. According to our plan some testees worked with the system for the limited period of time, equal for all of them. We designed an application which registered the knowledge level of the testees at the start and the end of the training course. The scheme of the work was similar for all of them, so it is possible to compare results of different participants.

There were two test groups, each of them worked with the software system once a week during a month. The testees were Russian and Bulgarian university and high-school students, most of them study linguistics. The Russian group worked with the native language and the Bulgarians learn Russian at the university.

Before the first evaluation we gave the participants basic information about LFs and the help material on LF presented in the test. During the test all answers (both correct and new to the system) of the testees were saved by our application. Because of that it is possible to trace back their work with the system step by step in details. Analyzing their answers which differ from the dictionary variants can give us the list of the most frequent mistakes and possible some material on the new answer variants.

After the first evaluation the testees had a training period. As they worked with the system they were shown the number of points won and all correct answers to the questions. Finally we repeated the evaluation on the same material as at the start of the training.

4.1. The linguistic material of the experiment

As the linguistic material of the experiment, we used a small part of the Russian dictionary of lexical functions. This material – a set of argument words and LF values for a number of LFs - was divided in two unequal parts. The larger part of arguments and values (called *training data*) was offered to the testees during the training period and the other part (called *reference data*) was only used for evaluation. There were two versions of the test – one for participants with basic linguistic knowledge and advanced test for those who study linguistics for long.

Some information on the basic level of the material is presented in Table 1. There are the list of LFs, the number of arguments for each of them and the number of LF values corresponding to these arguments. As we have said, some LFs have several values corresponding to one argument. If a testee will give one correct answer for each question given he or she will won the **Normal level of points**. There is a possibility to give several variants of the answer for each question, but after two correct answers the system shows the next question. So the testee can surpass the Normal level of points up to the **Accessible maximum of points**. The **General maximum of points** is a potential value which characterises the linguistic material; it couldn't be reached during the evaluation.

Table 1: The content of the basic level of the test.

N	Function name	LF difficul ty level	Number of arguments	Total number of values	Normal level of points	Accessible maximum of points	General maximum of points
1	Magn	1	47	133	47	119	219
2	Anti	1	26	31	26	34	36
3	Antimagn	1	21	38	21	43	55
4	Incep	1	7	9	7	9	11
5	Fin	1	10	20	10	24	30
6	Degrad	1	9	13	9	17	17
7	Sing	1	16	24	16	30	32
8	Mult	1	22	27	22	32	32
9	Equip	1	7	12	7	13	17
10	Gener	1	44	54	44	62	64
11	Loc	1	11	18	11	21	25
	Total		220	379	220	404	538

As we can see from the Table 1, basic level of the test contains 11 LF of the first difficulty level, it has total 220 arguments and 379 LF values, corresponding to these arguments. During the test the participants didn't see if their answer is correct or not. Also, they didn't see their final result.

Advanced level of the test consists of 22 LF – 11 LFs of the first difficulty level (Magn, Anti, AntiMagn, Incep, Fin, Degrad, Sing, Mult, Equip, Gener, Loc – same as in the basic level variant), 9 LFs of the second difficulty level (Bon, AntiBon, Caus, Func0, Func1, Func2, Labor, Oper1, Oper2) and 2 LFs of the third difficulty level (REAL1 and REAL1-M). The total number of arguments is 230, the number of values corresponding to them is 372. The **Normal level of points** is 369, the **Accessible maximum of points** is 614. The **General number of points** is 729. These numbers are much greater than numbers in the basic level of the test because participants are given more points for answers to the questions on difficult LFs.

4.2. General training results

We needed to have a mark of every test result to compare results of the testees and to calculate the value of their progress. The general idea of our software system is to encourage all correct answers; we do not reduce the mark for incorrect variants. So, traditional *f-value* which takes into account both *precision* and *recall* of the answers given is not the best mark for us. We decided to use *precision* as the basis of the mark – to see the part of normal level of points won by the testee, calculated for every LF and summarized with account of LF's weight in the whole test. We calculated the *average weighted mark* of every testee result at the start and the end of the training course, they are shown in Table 2.

The first results of our testees show that there is a gap in knowledge of idioms and collocations even in the group working with the native language – it's average mark is only 61,39%. The average mark of the testees who study Russian is expectedly lower and comes up to 51,45%. The total mark of all the participants for the first test is 55.87%.

Average mark for the final test of testees working with the native language is 98,98%, of those who study Russian is 89,34%. Average result of all the testees is 93,2%.

Table 2: Training results on both training and reference data.

Testee name	Testee group (N – native language, F – foreign language)	Test type (B – basic level, A – advanced level)	Average weighted mark, start of the course, % of normal mark	Average weighted mark, end of the course, % of normal mark	Value of the progress, %
Kira	N	B	86,82	130,00	43,18
Soroka	N	B	41,36	113,64	72,27
Nelya	N	B	85,00	125,45	40,45
Lena	N	A	51,76	53,66	1,90
Katya	N	A	63,41	80,22	16,80
Valja	N	B	43,64	90,91	47,27
Victorson	F	B	44,55	64,55	20,00
Sonya	F	B	51,36	41,82	-9,55
Darja	F	B	34,55	42,27	7,73
Inna	F	B	57,27	102,73	45,45
Ludmila	F	B	70,45	141,36	70,91
Masha	F	B	76,36	134,55	58,18
Nikolaj	F	B	40,91	67,27	26,36
Nina	F	B	64,55	138,64	74,09
Vanja	F	B	57,27	70,91	13,64

As we see, the final result of most participants surpasses their mark at the start of the course. The value of the progress varies from -9,58% to 72,28%, the average value is 35,25%. Negative value of the progress is registered only for one testee. It should be mentioned that despite a difference in average marks between the group of participants working with the native language and the group of participants who study Russian, the average progress of the first group (36,98%) is practically equal to the result of the second group. Minimum values of the progress correspond to three testees who study Russian and two testees who worked with the advanced level of the test. This fact is quite natural because the training time was rather small and the task of these participants was the most difficult.

4.3. Training results on the reference data

Training results on the *reference data* are shown in Table 3. First results on this material are very similar to the general results of the testees at the beginning of the training course. But the results of the final test are much more interesting.

Table 3: Training results on reference data.

Testee name	Testee group (N – native language, F – foreign language)	Arguments, start of the course, %	Arguments, end of the course, %	Value of the progress, arguments, %	LF values, start of the course, %	LF values, end of the course, %	Value of the progress, LF values, %
Kira	N	77,78	88,89	11,11	64,71	76,47	11,76
Soroka	N	50,00	80,56	30,56	35,29	74,51	39,22
Nelya	N	80,56	80,56	0,00	64,71	70,59	5,88
Lena	N	47,46	54,24	6,78	35,29	41,18	5,88
Katya	N	72,88	71,19	-1,69	54,12	58,82	4,71
Valja	N	58,33	69,44	11,11	47,06	54,90	7,84
Victorson	F	44,44	50,00	5,56	33,33	35,29	1,96
Sonya	F	66,67	52,78	-13,89	49,02	37,25	-11,76
Darja	F	44,44	50,00	5,56	35,29	37,25	1,96
Inna	F	52,78	69,44	16,67	39,22	54,90	15,69
Ludmila	F	72,22	75,00	2,78	60,78	62,75	1,96
Masha	F	72,22	83,33	11,11	58,82	68,63	9,80
Nikolaj	F	61,11	55,56	-5,56	43,14	49,02	5,88
Nina	F	63,89	80,56	16,67	54,90	64,71	9,80
Vanja	F	58,33	55,56	-2,78	45,10	39,22	-5,88

As we see, the number of correct values after the course is greater for the overwhelming majority of the participants, and the number of arguments, for which at least one correct value was given increased for the most of the testees. The average progress over the arguments is 6,26%, over the values 6,98%. These numbers are much lower than the progress over the whole training data. Also we can mention that results of the participants working with the native language (9,64% over the arguments and 12,55% over the values) are considerably higher than results of the group of participants who study Russian (4,01% over the arguments and 3,27% over the values).

The difference between the value of the general progress and the value of the progress on the reference data is the result of influence of some factor or several factors which distinguish work with these two parts of the test. There is the only one such factor, it consists in mastering by the participants the proposed material during the training, which doesn't affect the reference data. Progress over the reference data is possible only in case of mastering by the testees the idea of LF. A considerable fluctuation in progress values can be explained by the different level of understanding of the LF apparatus by different participants. The group of the testees who study Russian was in the most difficult situation, and that was the reason of distinction of their progress value from the average progress value of the group working with the native language.

4.4. Mark components – baseline and extra points

The total number of points won may comprise of two parts – baseline points and extra points. Baseline points are given to the testee for the first correct answer to the question and extra points are given for additional correct answers. Extra points are greater than the baseline points because it is more difficult to the testee to get them. So, there can be a strategy to answer as many questions as possible or to give as many answers as possible to separate questions.

It was interesting what part of the final progress value corresponds to the baseline points and what part corresponds to the extra points. We have divided the mark of the testees and have calculated its components separately. The average value of the progress is 35,25% and comprises of 16,79% of baseline points and 18,46% of extra points. We can see that extra points have the leading part of the total result.

It is also interesting that the value of the progress in baseline points is practically equal for the group of the testees working with the native language (16,89%) and for the group who study Russian (16,72%), but the value of the progress in extra points differs and makes up 20,09% for the first group and 17,37% for the second group. It can be explained by that fact that it was more difficult for participants who study Russian to master alternative LF values during the same training period than for participants who work with the native language. At the same time both groups showed equal results in mastering the baseline material.

5. Conclusions

The results of the training experiment are positive – most participants who passed the final test improved their mark for the whole test material (both for training data and reference data). This means that during the training the user masters the idea of lexical functions while simultaneously learning the dictionary content.

Acknowledgements We thank Ju. D. Apresjan, I. M. Boguslavsky, L. L. Iomdin, L. L. Tsinman, V. G. Sizov and L. G. Kreidlin very much for their help and great moral support at all stages of the work.

References

Boguslavsky et al. (2006) – Boguslavsky I., Barrios Rodrigues M., Diachenko P. CALLEX-ESP: a software system for learning Spanish lexicon and collocations. In Current Developments in Technology-Assisted Education (2006), Vol. 1, 2006, pp. 22--26.

Diachenko, P. (2006). Lexical functions in learning the lexicon. In Current Developments in Technology-Assisted Education (2006), Vol. 1, 2006, pp. 538--542.

Mel’čuk, I. A. (1974). Opyt teorii lingvisticheskix modelej "Smysl ⇔ Tekst" [A Theory of Meaning⇔Text Linguistic Models]. Moscow, Nauka, 1974, 314 p.

Apresjan et al. (2002) – Apresjan Ju. D., Boguslavsky I. M., Iomdin L. L., Tsinman L. L. Lexical Functions in NLP: Possible Uses. In: Computational Linguistics for the New Millennium: Divergence or Synergy? Proceedings of the International Symposium held at the Ruprecht-Karls-Universität Heidelberg, 21-22 July 2000. Manfred Klenner / Henriëtte Visser (eds.) Frankfurt am Main, 2002, pp. 55--72.

Apresjan, Ju. D., Tsinman, L. L. (2002). Formalnaja model perifrazirovanija predlozhenij dlja sistem pererabotki tekstov na estestvennyx jazykax [Formal model of sentence paraphrasing for NLP systems]. Moscow, Russkij jazyk v nauchnom osveshchenii, 2002, No. 2(4), pp. 102--146.