*Psychology of Language and Communication 2009, Vol. 13, No. 2*
DOI: 10.2478/v10057-009-0008-3

SOLOMIYA BUK
Ivan Franko National University of Lviv

# LEXICAL BASE AS A COMPRESSED LANGUAGE MODEL OF THE WORLD (ON MATERIAL FROM THE UKRAINIAN LANGUAGE)

In the article the fact is verified that the list of words selected by formal statistical methods (frequency and functional genre unrestrictedness) is not a conglomerate of non-related words. It creates a system of interrelated items and it can be named the "lexical base of language". This selected list of words covers all the spheres of human activities. To verify this statement the invariant synoptical scheme common for ideographic dictionaries of different languages was determined.

*Key words*: lexicography, linguistics statistics, lexical frequency, frequency dictionaries, functional genre, lexical base of language, Ukrainian language

## The selection principles of the Ukrainian language lexical base

In Ukrainian linguistic studies, dealing with modern lexical stratification, researchers investigate the lexical groups differing stylistically, by time and by territory, or by the environment of their functioning. The word's stratum with the highest usage and, according to V. Moskovič (Moskovič, 1969, p. 23-51), respectively, with the highest information density and the importance for text understanding, was not the single research object. Such a lexical base separation, its detailed analysis in terms of word composition and in terms of its classification into the paradigmatic groups can demonstrate the answer to the question about the language system. The paradigmatic group selection on this base, ascertaining different semantic relations between those groups, observing its semantic

Address for correspondence: Solomiya Buk, Department for General Linguistics, Ivan Franko National University of Lviv, 1 Universytetska St., Lviv, UA-79000, Ukraine. E-mail: solomija@gmail.com

Table 1. Comparison of frequency dictionaries in different languages

| Functional genre | Dictionary | | | | | |
|---|---|---|---|---|---|---|
| | French, % | Finnish, % | Slovak, % | Polish, % | Russian (Zasorina, 1977), % | Russian (Štejnfel'dt, 1963), % |
| belles-lettres | 20 | 11.5 | 30.2 | 20 | 25 | 12.5 |
| essay | 20 | | | | | |
| drama | 20 | | | 20 | 25 | 12.5 |
| poetry | | | 13.2 | | | |
| dialogue | | | 10.5 | | | |
| radio-program | | 9.2 | | | | 25 |
| journalistic | 20 | 26 | 14.6 | 40 | 25 | 25 |
| scientific | | | 31.5 | 20 | 25 | |
| literature for children | | | | | | 25 |
| different | | 43.3 | | | | |
| common corpus, word occurrences | 500 000 | 400 000 | 1 000 000 | 500 000 | 1 000 000 | 400 000 |

description in one language explanatory dictionaries will make easier the work on their adequate semantization.

The lexical base separation has a real theoretical foundation: one can consider the existence of the kernel vocabulary in any language as one of the universal features in human lexicon organization" (Serebrennikov et al., p. 121).

Practically all developed languages have such a lexical base, for example, English (Elridge, 1911; Thorndike, 1931; Ogden, 1937; West, 1953; Palmer, 1968), German (Hauch, 1931), Spanish (Keniston, 1929), French (Gougenheim et al., 1956), Polish (Kurzowa & Zgółkowa, 1992), Russian (Denisov, 1972; Morkovkin, 1984), etc. A comparison between lexical bases of different languages also exists (Eaton 1934). Practically all these authors (except Ogden, 1937) select their lists using statistical criteria from the frequency dictionaries, and some of them take into consideration a word's occurrence in different types of text. Taking into account previous experience, we developed our own techniques of language base selection.

As far as we know, there is no special research devoted to the quantitative correlation of the functional genres in the daily speech of an average person. There are many controversies in the point of proportioning and choosing the whole language frequency dictionary size in practice. Large frequency dictionaries are built on the basis of different proportions of genres. From this point of view we try to compare some frequency dictionaries of French (Juilland et al., 1970), Finnish (the information for Finnish language is from Tuldava, 1987, p. 56), Slovak (Mistrík, 1969), Polish (Saloni, 1990), and Russian (Zasorina, 1977, Štejnfel'dt, 1963) languages. The results are shown in Table 1.

As can be seen from Table 1, all the dictionaries take into account belles-lettres and journalistic genres, four of them (French, Polish and both Russian) consider drama as the equivalent of spoken language, three of them (Slovak, Polish, and Russian in Zasorina) take scientific texts into consideration. The fact of official genre lack attracts some attention. Certainly, it is somewhat presented in newspaper and magazine language, but cannot be confined to it. In order to select the lexical base, we decided to compare frequency dictionaries of five functional genres (due to the standard classification): belles-lettres, journalistic, colloquial (spoken language), scientific and official genres.

For the Ukrainian language, there are only two frequency dictionaries: belles-lettres (Perebijnìs, 1981) and journalistic (Darčuk & Grâznuhina, 1996); the principles of their construction are quite similar. Three others (colloquial, scientific, and official) were prepared by the author of this article (Buk, 2003a, 2003b).

Aiming all the functional style corpora under consideration to be equivalent, we used the corpus size of 300 000 word occurrences for each of three our dictionaries, according to the corpus size of the journalistic genre frequency dictionary.

For the further appropriateness of those frequency dictionaries comparing, their building principles were equal as described in Darčuk & Grâznuhina (1996). Our original frequency dictionary comparison (which takes into account statistical

Table 2. Comparison of word frequencies in five dictionaries (belles-lettres, colloquial, journalistic, scientific, and official genres)

| Word | Genres | | | | | |
|---|---|---|---|---|---|---|
| | belles- lettres | colloquial | journalistic | scientific | official | sum |
| *новий* | 262 | 155 | 495 | 434 | 179 | 1525 |
| *тому* | 206 | 444 | 296 | 379 | 171 | 1496 |
| *організація* | 14 | 12 | 460 | 205 | 745 | 1436 |
| *можна* | 262 | 419 | 353 | 370 | 32 | 1436 |
| *слово* | 445 | 337 | 415 | 208 | 23 | 1428 |
| *процес* | | 20 | 152 | 1111 | 136 | 1419 |
| *питання* | 43 | 74 | 521 | 283 | 477 | 1398 |
| *увесь* | 254 | 455 | 403 | 173 | 110 | 1395 |
| *місце* | 223 | 202 | 330 | 240 | 380 | 1375 |
| *український* | 56 | 14 | 703 | 314 | 262 | 1349 |

methods and world experience analyses) is described in Buk (2003c). In particular, it takes into account the text coverage analysis.

A special program was written for such a frequency dictionary comparison. It brings together all the dictionary words in one (first) column named "word", in the next columns (they are indexed by the numbers of five frequency dictionaries) every word frequency is fixed. The last column shows the word sum for all the dictionaries (see Table 2).

The common lexical base size is 1389 words.

## The methodology of revealing for the conceptual model of the world

It can be a very important result if the selected list of words covers all the spheres of human activity. To verify this statement, it would be good to have a conceptual or language model of the world. The conceptual model of the world, in our opinion, can be brought to light by comparing the ideographic dictionaries in different languages. Our hypothesis is the following: there is an invariant synoptical scheme irrespective of language in all ideographic dictionaries. It is caused by the fact that human knowledge has a systematic nature, and language (in particular, the lexical composition) is its main vehicle, so they should be a similar system.

For this purpose we tried to collate the ideographic dictionaries synoptical schemes of English (Roget, 1977), German (Hallig & Wartburg, 1963; Meier, 1964; Dornseiff, 1970), Spanish (Casares, 1959), Czech (Haller, 1974), Russian (Morkovkin,

1984) and Ukrainian (Sokolovs′ka, 2002). First, we review very shortly those schemes without detailed description of their positive or negative aspects, aiming to show their general world-view differences. It is important to note that these are not linguistic, but rather logical, classification schemes of concepts.

*Roget's International Thesaurus* (Roget, 1977) divided the English vocabulary as a first step into eight groups with subdivisions:

I "Abstract relations" (existence, relation, quantity, order, number, time, etc.);
II "Space" (dimensions, structure, form, motion);
III "Physics" (heat, light, electricity and electronics, mechanics, etc.);
IV "Matter" (inorganic matter, organic matter);
V "Sensation" (touch, taste, smell, sight, hearing, sound);
VI "Intellect" (intellectual faculties and processes, state of mind, communication of ideas);
VII "Volition" (condition, voluntary action, authority and control, support and opposition, possessive relations);
VIII "Affection" (personal affections, sympathetic affections, morality, and religion).

F. Dornseiff (1970) divided the German vocabulary into 20 groups with the next smaller subdivision. The first and the second one cover the nature, which is understood here very widely: from cosmos, meteors, inorganic world through plants and animals to the human body. The next six groups include the abstract and a priori concepts ("space", "size", "existence", "time", etc). The next four groups consist of human psychological characteristics: "wishes and actions", "sensation", "feeling, affects, feature of character", "thought"). The words of the four last groups describe social relations and cultural phenomenon.

Another German language division was proposed by R. Hallig and W. von Wartburg (1963). They divided the universe into three main spheres: "universe", "human being", "human being and universe". Each of these spheres covers several conceptual fields, and in the sum there are ten big complex fields ("heaven and heavenly bodies", "earth", "plant world", "animal world", "man as an alive being", "soul and mind", "man as a social being", "social organization and social institutions", "a priori", "science and technique"). Those fields have the next division.

The similar scheme lies in the basis of *Česky slovník věcný a synonymický* (Haller, 1974). The authors write in the preface that they depart from the R. Hallig and W. von Wartburg dictionary only in the case where Czech material needed another classification (Haller, 1974, p. V). In practice, the difference between the schemes of both dictionaries is a minimum.

H. Meier (1964) has done a statistically based synopsis. He has divided all the German vocabulary (11 million word occurrences) into 12 frequency zones: the first includes the most frequent words, the last includes the least frequently words. Û. Karaulov mentioned an interesting fact of the close result of two vocabulary classifications (by H. Meier and by R. Hallig and W. von Wartburg) obtained by different methods (Karaulov 1967, p. 254).

J. Casares (1959) built his Spanish language dictionary scheme with God in the center. After God "universe" follows divided into inorganic ("matter and energy", "physics and chemistry", "geography, astronomy, meteorology", "geology, mineralogy") and organic matter (plant and animals). The animal world includes both "animal" and "man", the last group consists of "individual" and "society" with the following subdivision of individual into the groups: "human as a living being", "human as an intellectual being" and "human as an agent of action", and the "society" divided into "communication, senses, thoughts", "social institutions", "work, service".

A. Markowski (1990) created the scheme of Polish language with the word "I" on the top and three main fields: "I in relations with myself" and "I in relations with others" (with relations with other people and other things). In the first field are: "I as a physical being" ("my body" and "something serving my body") and "I as a psychic being" ("my thought" and "something serving my thought"). In the second are: "I in relation with God" ("my belief" and "something serving my belief"), "I in relations with people" ("my attitude to others" and "something serving me and others").

V. Morkovkin (1984) proposed the hierarchic conceptual worldview of the Russian language with regard to teaching methodics. In the "universe" on the base of a dichotomous division he has divided conceptual spheres as follows: "abstract relations" – "material matter", "inorganic world" – "organic world", "plants" – "alive being", "unwise alive being" – "human being". In "Abstract relations" seven general groups are separated: "existence", "space", "time", "changing", "quantity", "quality" and "relations".

The Ukrainian scholar Ž. Sokolovs′ka (2002) has built a universal frame for any language (including Ukrainian) on gnoseological and ontological parameters. The gnoseological concepts (cognition categories) such as existence, space, time, movement, something separate, quality, quantity, relation, are in the vertical column of table and ontological concepts (existence spheres), such as nature, man, society, are in the horizontal line. There are the words in the square where the lines cross.

After collating the ideographic dictionaries synoptical schemes of six different languages (see above) we can see in their center the common invariant part as follows: nature, including the spheres from heaven to animals, human beings with the body and mental features, the relations between people in society, and independent categories like existence, space, time, movement, etc.

## Specifics of the semantic structure of the Ukrainian language lexical base

With the aim to find out what spheres of logically classified concepts are covered by our existent word list, we should classify this list itself. Although different nationalities use the same scientific conceptual instrument, some concepts can have no separate lexemes for its notation in some languages, for instance, English *blue* – Ukrainian *sìnij, golubij*; English *love,* Russian *lûbit΄* – Ukrainian *lûbiti*, *kohati* etc. So, it can not have an equal classification and we do not agree with R. Tokarski who equates the lexical and conceptual fields (Tokarski, 1984, p. 11). That is why we consider the semasiologic approaches to the vocabulary classification to be more natural for exact language classification because it is not fastened to the logical scheme for words but goes from word to concept.

Our technique of language base classification was the following: in the first stage, parts of speech (as the most general linguo-philosophic categories) were selected. There were nine of them: noun, verb, adjective, pronoun, proverb, numeral, preposition, conjunction and particle. No interjection was found. There is also no article in Ukrainian.

In the second stage, based on the common semantic features within parts of speech, the words were joined into small groups: synonymic rows, antonymic pairs, hypero-hyponymical, partial-holonomy ("meronymy" in Lyons' term (Malmkjær, 1991, p. 301) and conversion-based groups. Different group types were found in different parts of speech. Synonymic and antonymic rows were found in all of them: synonymic (*šlâh* 'way', *doroga* 'road'; *zahoditi* 'to enter', *vhoditi* 'come into'; *tâžkij* 'hard', *važkij* 'difficult'; *vìl΄nij* 'free', *nezaležnij* 'independent'; *zvičajno* 'obviously', *očevidno* 'evidently'; *bìlâ* 'nearly', *poruč* 'close (to)'; *os΄, ot* 'amplifier particle', etc.) and antonymic (*nadìâ* 'a hope' – *strah* 'a fear'; *zahoditi* 'to enter' – *vihoditi* 'to leave'; *holodnij* 'cold' – *garâčij* 'hot'; *švidko* 'fast' – *dovgo* 'long'; *do* 'to' – *vìd* 'from'; *tak* 'yes' – *nì* 'no'; *ŝe* 'yet' – *vže* 'already', etc.).

Hypero-hyponymical groups were found in the nouns, verbs and adjectives (*kìmnata* 'room' – *kabìnet* 'cabinet', *klas* 'class', *zal* 'hall'; *počuvati* 'to feel' – *lûbiti* 'to love'; *lûds΄kij* 'human' – *žìnočij* 'feminine', etc.).

Conversion-based group are found in nouns, verbs, prepositions, and interjections (*čolovìk* 'husband' – *družina* 'wife', some noun pairs of the model "*pričina* 'a cause; a reason' – *naslìdok* 'effect' ": *dati* 'to give' – *vzâti* 'to take', *sered* 'in the middle' – *navkolo* 'round'; *jâkŝo* 'if' – *to* 'then' etc.).

And the partial-holonomy groups were found only in nouns (*tìlo* 'body' – *golova* 'head', *ruka* 'hand', *noga* 'leg'; *ruka* 'hand' – *palec΄* 'finger', etc.)

Then, on the third stage depending on the specifics of the semantic value of each word (denotative- or significative-based) these small groups were joined into lexical-semantic or thematic groups. The verbs create the lexical-semantic groups only, but the noun, pronoun and adverb have the lexical-semantic as well as thematic groups. For example, the nouns with denotative-based lexical meaning of natural formation create the thematic group corresponding to it: *gora* 'mounting', *pole* 'field', *lìs* 'forest', *step* 'steppe', *more* 'see', *rìčka* 'river'. The nouns with significative-based lexical meaning of time create the lexical-semantic group: *čas* 'time', *rìk* 'year', *mìsâc΄* 'month', *tižden΄* 'week', *den΄* 'day', *godina* 'hour', *hvilina* 'minute', etc. The pronoun can be combined into lexical-semantic (e. g., "group of space": *korotkij* 'short', *visokij* 'high', *niz′kij* 'low', *glibokij* 'deep') and thematic groups (e. g., "group of production": 'production', *trudovij* 'working', *robočij* 'trade', *profesìjnij* 'professional', *tehnologìčnij* 'technological') and so on. The lexical-semantic groups of time, movement, relation, space, etc. were distinguished in all the parts of speech.

Based on these lexical-semantic groups in the case of verbs, the lexical-semantic fields of movement, state, relation and others were distinguished. For nouns, groups cannot be so strictly organized in such discrete fields. The most relevant differential features for noun meaning are: concrete / abstract. Within the concrete nouns the words were joined into animate / inanimate nature, human being and social relations. Within the abstract nouns the relevant feature was what kind of concept the word is connected with: a man, his work, mental or body characteristic, with nature or with abstract categories. We discovered the close situation in adjectives and in adverbs.

The last stage of lexical base classification is the crystallization of general lexical fields covering all parts of speech. There are fields of man, his body, mental features and mind, his work, individual relationships and attitude, social institutions and bureaucracy, animate and inanimate nature, general categories like time, space, existence, quality, quantity, and some others. As we can observe, the word fields are quite correlative with conceptual groups from the invariant base of all the ideographic dictionaries.

But there are some distinctive features. For example, we can see the general tendency of lexical base abstractness. It became apparent not only in the large number of abstract nouns, but in verbs general meaning as well. In many cases in the lexical base is only the verb (the most neutral) naming the whole field or group in the ideographic dictionary (*govoriti* 'to say' but not *šepelâviti* 'burr', *kričati* 'cry', *šepotìti* 'whisper', etc). There are large groups of words connecting with the norm (*tipovij* 'typical', *normal′nij* 'normal', *normativnij* 'normative', *vìdpovìdnij* 'corresponding', *zvičajnij* 'usual', *prirodnij* 'natural', *osoblivij* 'special', etc.), working process (*stadìâ* 'stage', *etap* 'phase', *metod* 'method', *sposìb* 'manner', *tehnologìâ* 'technology', *priom* 'technique', *režim* 'procedure', etc.), leading profession (*kerìvnictvo* 'leadership', *prezident* 'president', *direktor* 'director', *kerìvnik* 'chief', *zastupnik* 'deputy director', etc.). We should take note of absence of such groups as taste, sides of the world, seasons, days of the week. It is striking that there are *sìogodnì* 'today', *zavtra* 'tomorrow' but no *včora* 'yesterday'; there is *dorogij* 'expensive', but there is no *deševij* 'cheap'; there is *žìnočij* 'feminine' but no *čolovìčij* 'masculine', there are *garâčij* 'hot' and *holodnij* 'cold' but no *teplij* 'warm'.

At this stage we can only establish the existence or absence of some of the words with some meanings, but the explanation of this phenomenon can be done only after future research. A concomitant result of our analysis is the partial answer to the question "how language could be related to the world", considered by D. Geeraerts (Aszer, 1994, p. 3804).

In spite of some indicated discrepancy, the list of words selected via formal techniques using the criteria of frequency and functional unrestrictedness covers practically all the conceptual fields. From this point of view, this list, being the lexical base of the Ukrainian language, might be called the compressed model of the world.

## References

Aszer, R.E. (Ed.) (1994). *The encyclopedia of language and linguistics.* Vol. 7. Oxford: Pergamon Press.

Buk, S. (2003a). Častotnij slovnik naukovogo stilû sučasnoï ukraïns´koï movi. *Vìsnik Čerkas´kogo unìversitetu. Serìâ filologìčnì nauki*, 44, 90-96.

Buk, S. (2003b). Častotnij slovnik rozmovno-pobutovogo stilû sučasnoï ukraïns´koï movi. *Lìngvìstičnì studìï*, 11 (1), 266-271.

Buk, S. (2003c). Metodika vidìlennâ âdra leksičnogo mìnìmumu ukraïns´koï movi. *Mova ì kul´tura*, 4 (5/1), 25-31.

Casares, J. (Ed.) (1959). *Diccionario ideológico de la lengua española: desde la idea a la palabra, desde la palabra a la idea.* Barcelona: Gustavo Gili.

Darčuk, N.P. & Grâznuhina, T.A. (1996). Častotnij slovnìk sučasnoï ukraïns´koï publìcistiki. *Movoznavstvo*, 4/5, 15-19.

Denisov, P.N. (1972). *Leksičeskie minimumy russkogo âzyka.* Moskva: Moskovskij gosudarstvennyj universitet.

Dornseiff, F. (1970). *Der Deutsche Wortschatz nach Sachgruppen: Versuch eines Ordnungsschemas.* Berlin: de Gruyter.

Eaton, H. (1934). Comparative frequency list on the first thousand words in English, French, German, and Spanish. In A. Coleman (Ed.), *Experiments and studies in modern language teaching* (p. 244-279). Chicago: University of Chicago Press.

Elridge, R.C. (1911). *Six thousand common English words.* Buffalo: The Clement Press.

Gougenheim, G., Michea, R., Rivenc, P., & Sauvageot A. (Eds.) (1956). *L'élaboration du français élémentaire: étude sur l'établissement d'un vocabulaire et d'une grammaire de base.* Paris: Didier.

Haller, J. (Ed.) (1974). *Česky slovník věcný a synonymický.* Vol. 1-2. Praha: Státní Pedagogické Nakladatelství.

Hallig, R. & von Wartburg, W. (1963). *Begriffssystem als Grundlage für die Lexikographie.* Berlin: Akademie Verlag.

Hauch, E.A. (1931). *A German idiom list selected on the basis of frequency and range of occurrence.* New York: MacMillan.

Juilland, A., Brodin, D., & Davidovitch, C. (1970). *Frequency dictionary of French words.* The Hague: Mouton.

Karaulov, Û.N. (1976). *Obŝaâ i russkaâ ideografiâ.* Moskva: Nauka.

Keniston, H. (1929). *Spanish idiom list.* New York: MacMillan.

Kurzowa, Z. & Zgółkowa, H. (Eds.) (1992). Słownik-minimum języka polskiego: podręcznik do nauki języka polskiego dla szkół podstawowych i obcokrajowców. Poznań: SAWW.

Malmkjær, K. (Ed.) (1991). *The Linguistics Encyclopedia.* London: Routledge.

Markowski, A. (1990). *Leksyka wspólna różnym odmianom polszczyzny.* Vol. 1-2. Warszawa: Uniwersytet Warszawski.

Meier, H. (1964). *Deutsche Sprachstatistik.* Bd. 1. Hildesheim: G. Olms.

Mistrík, J. (1969). *Frekvencia slov v slovenčine.* Bratislava: SAV.

Morkovkin, V.V. (1984). *Leksičeskaâ osnova russkogo âzyka: Kompleksnyj učebnyj slovar´.* Moskva: Russkij âzyk.

Moskovič, V.A. (1969). *Statistika i semantika.* Moskva: Nauka.

Ogden, C. (1937). *Basic English. A general introduction with rules and grammar.* London: Kegan Paul, Trench, Trubner and Co.

Palmer, H.E. (1968). *The scientific study and teaching of languages.* Oxford: Oxford University Press.

Perebijnìs, V.I. (1981). *Častotnij slovnik sučasnoï ukraïns´koï hudožn´oï prozi.* Kiïv: Naukova dumka.

Roget, P.M. (1977). *Roget's International Thesaurus.* New York: HarperCollins.

Saloni, Z. (Ed.) (1990). *Słownik frekwencyjny polszczyzny współczesnej.* Vol. 1-2. Kraków: Uniwersytet Jagielloński.

Serebrennikov, B.A. & Kubrâkova, E.S (Eds.) (1988). *Rol´ čelovečeskogo faktora v âzyke. Âzyk i kartina mira.* Vol. 1. Moskva: Nauka.

Štejnfel´dt, È.A. (1963). *Častotnyj slovar´ sovremennogo russkogo literaturnogo âzyka. 2500 naibolee upotrebitel´nyh slov.* Tallinn: Progress.

Sokolovs´ka, Ž.P. (2002). Kartina svìtu ta ìêrarhìâ sem. *Movoznavstvo*, 6, 87-91.

Thorndike, E. (1931). *The teacher's word book.* New York: Columbia University Teachers College.

Tokarski, R. (1984). *Struktura pola znaczeniowego. Studium językoznawcze.* Warszawa: PWN.

Tuldava, Û.A. (1987). *Problemy i metody kvantitativnosistemnogo issledovaniâ leksiki.* Tallinn: Valgus.

West, M. (1953). *A General Service List of English Words, with semantic frequencies and a supplementary word-list for the writing of popular science and technology.* New York: Longman Dictionaries.

Zasorina, L.N. (Ed.) (1977). *Častotnyj slovar´ russkogo âzyka.* Moskva: Russkij âzyk.