Adam Pawłowski

QUANTITY STRUCTURE OF LEXICAL FIELDS AND THE HUMAN COGNITIVE PROCESSES[1]

Originally published as "Struktura ilościowa pól leksykalnych a procesy poznawcze człowieka," Studia Semiotyczne 27 (2010), 71–80. Translated by Agnieszka Ostaszewska.

INTRODUCTION

The notion of optimum coding is traditionally associated in linguistics with a binary record of the phonetic level units and/or other equivalent symbols, e.g. letters (Jassem 1974: 239-297; Herdan 1966: 259-303; Hammerl, Sambor 1990: 418-420). Its purpose is to demonstrate that variable frequencies of occurrence of the examined symbols may reduce the length of the coded message, i.e. indirectly reduce the time of its processing (understanding, reproduction and/or sending). A similar technique may, however, also be applied to the subsystems of the language. In particular, with the use of this method, one may code lexical fields, treated here as subsets of the lexical subsystem.

Comparing, on the basis of corpus and dictionary data, lexical fields in various languages, one is able to note that the frequencies of lexemes in each field are uneven and distributed in a decreasing order in a quite characteristic and predictable way (a non-monotonic decreasing curve of the shape resembling a negative-exponent power function, Fig. 1). This phenomenon is present in various languages, probably in the case of all lexical fields, but also on various levels of generality, since it pertains to the quantity structure of the entire vocabulary, and not only the subsets thereof. On the surface, the explanation seems simple: namely, it is possible

[1]The analyses presented in this paper are an elaboration of the empirical analyses contained in Pawłowski 2007.

to refer to the categories of COGNITIVE REALISM and assume that the diversification of the frequencies of lexemes reflects the quantity structure of actual designations of the respective notions. This would mean for example that lexemes *black, white, red* etc. have the highest frequencies, since they indicate the colours which are most often found in the environment of a human being (an analogous reasoning may be carried out with reference to lexical fields representing other fields of reality). However, this explanation is insufficient. Although there is a certain relation between the human environment and the quantity properties of vocabulary, explanation of uneven distributions of frequencies in the lexical fields requires a reference to the basic epistemological categories (constructionism, cognitive realism, apriorism, and aposteriorism), as well as to the knowledge of neuro-linguistic processes taking place during the recollection, reproduction and transmitting of linguistic information by a human being.[2]

THE CONCEPT OF A LEXICAL FIELD

The term lexical (word) field[3] has entered the linguistic conceptual apparatus thanks to the works of J. Trier, W. Porzig, L. Weisgerber, E. Coseriu and other, mainly German, scholars, active in the first half of the 20^{th} century (Trier 1973a, 1973b; Weisgerber 1950; Porzig 1957; Coseriu 1975). In Polish scientific literature this view was popularised i.a. by Walery Pisarek (1967), Danuta Buttler (1967) and Ryszard Tokarski (1984, 2006). It is worth adding that the difference between the scopes of the sometimes interchangeably used names of *the lexical field* and *the semantic field* is relatively small: the notion of the lexical field is defined in the semasiological perspective, which is characterised by the primary nature of the network of names (signs) with respect to the network of notions and the set of designations; the notion of the semantic field is, on the other hand, defined in the onomasiological perspective, where the supreme role is played by the relations between the designations and not their names (signs).

A lexical field may be defined as a set of lexemes having common semantic properties. An example may be the lexical fields of the names of colours, animals, plants, vehicles, etc. The most distinct elements of this

[2]Literature devoted to the examination of linguistic data coding processes in the psychological perspective is quite extensive. One of the latest works on the subject is Michael Fortescue's *A Neural Network Model of Lexical Organisation* (Fortescue 2009; cf. also Johnson 1978).

[3]The term *word field* is a loan translation of the German name *Wortfeld* and shall be treated as a synonym of the name *lexical field*.

set may be recognized on the basis of the fact that they are combined by a relation of hyponymy with the same hypernym. Less distinct elements, i.e. such elements whose appurtenance to a given lexical field is disputable, may be recognized on the basis of the fact that they are in a relation of broadly understood meronymy with respect to the basic lexemes, and do not have to demonstrate the same morphosyntactic properties (for example the lexemes: *to paint, light* and *shadow* in relation to the "core" of the lexical field of names of colours).

An issue more important from the definition of the lexical field itself however, is the idea to treat vocabulary as a large system, composed of smaller, coherent subsystems, and not as an amorphous set of independent, isolated units. This idea is, of course, present to a varying degree, in the assumptions of many theories of linguistics and other disciplines related to linguistics. And thus, close to the notion of the lexical field are the notion of Ch. Fillmore's SEMANTIC FRAME (Fillmore, Atkins 1992), psycholinguistic notion of BEHAVIOURAL SCRIPT, the notion of ONTOLOGY in language engineering and AI research, the notion of SYNSET in Wordnet research (cf. Miller 1998, Piasecki et. al 2009) and the notion of SEMANTIC NEST in lexicology and psycholinguistics (Sambor 1997; Sambor, Hammerl 1991; Łobacz, Mikołajczak-Matyja 2002). In this paper it has been assumed that not only the vocabulary, but also THE REPRESENTATION OF KNOWLEDGE IN THE HUMAN MIND IS OF SYSTEMIC CHARACTER, AND ONE OF THE METHODS THAT MAKES IT POSSIBLE TO DISCOVER HUMAN COGNITIVE SCHEMATA IS TO EXAMINE THEIR EXTERNAL MANIFESTATIONS, FOR EXAMPLE THE LEXICAL FIELDS. Such research may in consequence facilitate the choice between two competitive epistemological approaches, which one may consider to be cognitive realism and constructivism.

LEXICAL FIELDS STRUCTURE MODELLING METHODS

From a mathematical point of view, distribution of the frequencies of lexemes may be described by many methods. One of the methods which is often applied is function estimation, constituting assumingly a theoretic model of the described phenomenon. This type of modelling was propagated by the German school of quantitative studies (cf. Altmann 2000; Köhler et al. 2005). Function models of this kind have many advantages, they show the co-dependency between the variables, as well as provide for a prediction of the properties of texts of a given kind. Their disadvantage are their small explanatory power, and therefore lack the possibility to explain the essence

of the phenomenon, its sources and consequences. Approaching this issue in a minimalist way, it is of course possible to assume that the reason for the variability in the value of function *f(x)* modelling the studied phenomenon, are the changes of the value of parameter x, it is also possible to show the dynamics of these changes. A model treated in this way is not an explanation, however, but only a mathematical, formalized mapping of a certain fragment of physical or abstract reality.

In order to avoid limitations, this paper employs an approach, whose objective and main idea is to search for the causes of the phenomenon, and not only the mapping of its internal dynamics. It has been assumed that a subset of the lexicon, constituting a lexical field, may be represented in the memory of a person as a binary sequence, and therefore, its model should also be based on a two-value scale. This approach is consistent with the current knowledge of neurological processes, since zero and one in the mathematic model correspond to the states of activity and non-activity of a neuron. The neuron activation process, taking place at the synapse, consists of adjusting the amount of the neurotransmitter, i.e. of the substance separating the ending of the axon of the neuron delivering information from the receiving neuron, which makes it possible to transmit an electrical impulse between the neurons. It may be added that in the artificial neural network theories, the neuron activation process is modelled by the so-called binary-type threshold function (Tadeusiewicz 2000: 4-17; Rutkowska et al. 1999: 18-21).

Two methods have been applied in order to code binary sequences corresponding to single lexemes:

– simple coding, based on the principle that binary sequences corresponding to particular lexemes are of the same length;

– optimal coding, based on the principle that the length of the binary sequence depends on the frequency of a lexeme's occurrence, whereby frequent lexemes are coded with the use of shorter sequences and non-frequent lexemes are coded with the use of longer sequences (the so-called Huffman coding).[4]

Having performed the coding, we have compared the average lengths of the binary sequences, corresponding to the lexemes belonging to the lexical field of colours, obtained by both methods. The obtained results

[4]Optimal coding technique, used most often by data compression, is based on a quite simple algorithm, which has been described in the literature from the field of linguistics, study of information and computer science (Meyer-Eppler 1959; Hammerl, Sambor 1990: 415-423), as well as on various all-accessible WebPages (examples of such descriptions may be found at: `http://en.wikipedia.org/wiki/Huffman_coding`, `http://www.compressconsult.com/huffman`, `http://www.quant-dec.com/Articles/steganography/huffman.ht`).

have been subjected to analysis and interpretation and an account taken of evolutionary aspects of the increase in the efficiency of human brain cognitive processes in the course of phylogenesis. By formulating conclusions, it has been assumed without proof that processes and systems of communication in the world of living organisms are governed by two basic principles. The first is the principle of the least effort. It stipulates that each organism strives at minimising the amount of energy invested in the process of generating, understanding, remembering and sending information. According to the second principle, the communication systems are relatively autonomic and are subject to self-regulatory processes. One of the consequences of the operation of these principles is the common phenomenon of the abbreviation of forms of high frequencies, modelled i.e. by Zipf's law.

RESEARCH RESULTS

In the first part of this research a histogram was prepared of the average frequencies of basic colour names in ten Indo-European languages, based on a representative five-million sample (in each language there were on average 500 000 test words).[5] The absolute values were changed into a percentage share of particular colour names in the entire lexical field and presented in decreasing order (Fig. 1). The obtained result has a distribution typical for most lexical fields, which may be observed in the vocabulary structure of a single text of relevant length,[6] a collection of texts, as well as the entire vocabulary of a given language. This distribution has not been subject to modelling, however, one may suspect that a non-monotonically decreasing function (e.g. a power function or an exponential function) would yield good results in this case. The data was then coded using Huffman's method.

[5]A detailed description of the corpus of the texts is contained in the works Pawłowski 2003 and 2007.

[6]Balance, whereby increase of the length of the sample does not materially affect the value of the measured parameters, is considered to be the most reliable criterion specifying the text volume in quantitative language study.

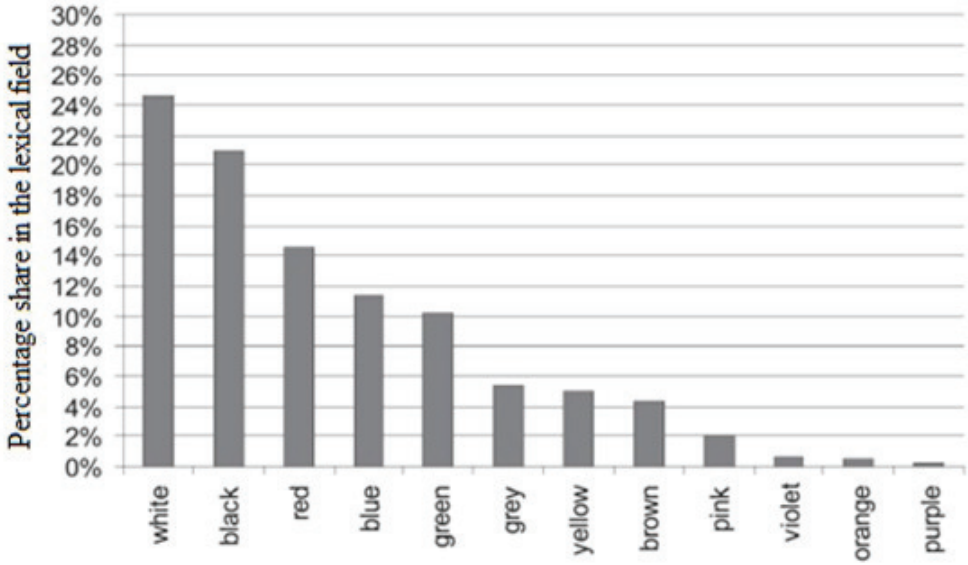

Fig. 1. Share of average colour names frequencies in the lexical field on the basis of a multi-lingual corpus of texts (cf. Pawłowski 2003 and 2007).

In accordance with the expectations, the average length of the sequences of bytes coded with the optimal method proved to be smaller than the average length of the sequences obtained

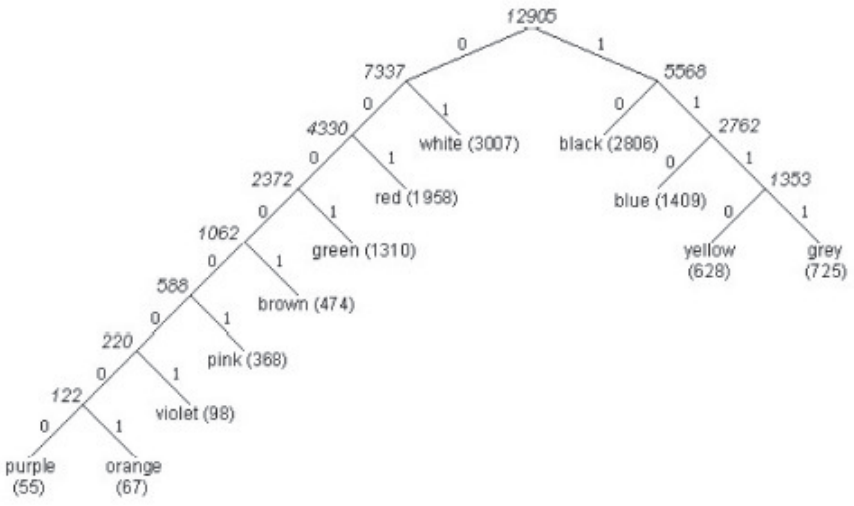

Fig. 2. Huffman codes values ascribed to the elements of the lexical field of colours

C	F	p	p.c.	H.c.	L
white	3007	0.233	1111	10	2
black	2806	0.217	1110	01	2
red	1958	0.152	1101	100	3
blue	1409	0.109	1100	011	3
green	1310	0.102	1011	1000	4
grey	725	0.056	1010	1111	4
yellow	628	0.049	1001	0111	4
brown	474	0.037	1000	10000	5
pink	368	0.029	0111	100000	6
violet	98	0.008	0110	1000000	7
orange	67	0.005	0101	10000000	8
purple	55	0.004	0100	00000000	8

Tab. 1. Binary coding of the elements of the lexical field of colours

Designations:
C name of the colour
F frequency of the use of terms corresponding to C in the corpus
p empirical probability of C appearing in the corpus
p.c. proportional coding (sequences of even length)
H.c. optimal coding (sequences of varying length)
L length of optimally coded sequence

in the course of even coding. The length of the sequence by even coding for the lexical field of the colours was permanent and was equal to 4 bytes of information, whereas by optimum coding it dropped to 2.97 bytes (Tab. 1, Fig. 1). This means that the increase in the effectiveness of processing the information coded by Huffman's method was equal to about 25%, since this is the value by which the average length of a "binary word" was reduced, and therefore by which the average time of decoding, recording or sending it was reduced. The notion of information is wide, of course (in the general sense it means every stimulus increasing the organism's knowledge of its environment). In this context, however, information should be associated with processing (coding, decoding or sending) of a stimulus corresponding to one notion or lexeme.

CONCLUSIONS

The cause-and-effect reasoning, whose objective is to explain the common phenomenon of uneven distributions of frequencies of lexical units in

communication systems, is based, as already mentioned above, on the principle of the least effort. In connection with the self-regulatory mechanism, these are the principle results in establishment of a state of balance between two conflicting forces, which are, on one hand the human communication efficiency and orientation in the environment, and on the other — limited capabilities of the human brain to register and process information. The maximisation of the first parameter, i.e. the best possible orientation of a human being in the environment, would require processing, in real time, of a practically unlimited number of various stimuli, constantly received by the preceptors. However, such a task exceeds the capabilities of the human brain. This is why probably in the course of the phylogenetic adaptation processes there have developed internal cognitive mechanisms, which in a way pump the stream of stimuli into ready, simplified schemata (the model of such schema has been presented in Fig. 1 and 2). The HYPOTHESIS OF COMPROMISE between the tendency to maximise the quantity of analysed information and the limited capabilities of the human brain to process information, which has been presented and initially verified herein, is a very good starting point, leading to formulation of generalisations.

The first conclusion that comes to mind after analysis of the data is the assumption that human cognitive processes are of MODERATELY APRIORICAL character. This means that the representation of knowledge contained in the human brain is determined not by external stimuli but by the structure of the memory itself. It forces data categorisation based on uneven distributions, composed of ca. seven or eight units of decreasing frequencies and of a large number of low frequencies ("the curve's tail"). It may be said that people perceive reality in this manner and nothing else, since the brain would not stand the intensity of the cognitive process, in which every element of the world would be categorized in accordance with its physical properties, and the units belonging to lexical fields would have similar values in the discourses. Distinguishing for example n perceptively separate colours (n may vary from several hundred to several thousand, depending on individual features), a human being reduces this number in his representation of knowledge to several dominant terms, described as basic colour terms (Kay, Maffi 1999; Pawłowski 2006). Although loss of information takes place, it is compensated by the greater speed of processing smaller numbers of categories, which finally increases the human being's orientation in his environment. This phenomenon is of course multiplied by reference to all linguistically categorised elements of experience.

This conclusion does not mean, however, that the verbalised represen-

tation of knowledge is entirely separate from experience. Perception, and therefore indirectly the human being's environment, is decisive for which categories are to be put in particular places of the schema analogous to the one presented in Fig. 1. For example, the primal experience of light, darkness, blood and fire results in the fact that colours corresponding to these prototypical phenomena or designations are present, in all languages which were linguistically examined, at the first three places in the schema. Also of importance here are the physiological properties of the human eye, which provide for easier recognition of certain colours. However, the mere scheme of decreasing frequencies of subsequent lexemes, resulting in a subjective conviction of the language users of varying "importance" or "prototypicality" of particular colours, is only a result of the limitations imposed by the human brain. Since an analogous reasoning is possible to be carried out with respect to lexemes constituting other lexical fields, the conclusions presented here should be considered relevant for the entirety of the human cognitive processes.

In order to find additional confirmation of this conclusion, one might design an experiment consisting therein that a group of people is located in an isolated, yet observable, environment (a kind of "scientific Big Brother"), containing even distribution of perception stimuli of a certain type. It needs to be expected that as a result of self-regulation and optimisation of the cognitive process, the linguistic representation of this balanced group of stimuli, corresponding to a certain lexical field, will not be even, but will adjust to the schema built in the human psyche, presented in Fig. 1. The relation of this structure with such constructs as Universal Grammar or *Lingua Mentalis* remains an open question, however (the conducted research allows only to state with high probability that such relation does exist).

The second conclusion is of self-referential character, and its consequences may prove auto-destructive. Since the representation of knowledge in the human brain has such a large autonomy in relation to the sensually perceived reality, then also perhaps the entire human knowledge, to which the cognitive activities of the human mind lead, should be recognized as a construct only loosely connected to the reality (self-reference means here recognition of this paper as being an element of the scientific discourse). Such a conclusion would be consistent with the standpoint of radical constructivism, which stipulates that "[...] human beings, due to the construction of their nervous system, do not have cognitive access to reality. The human nervous system is autopoietic and self-referential, semantically and operationally closed. All we can do is construct reality" (Graszewicz, Lewiński 2007: 206). The conducted

research does not, however, justify drawing such extreme conclusions. It has only been demonstrated that "access to reality" is strongly distorted by human adaptation mechanisms, optimising the cognitive process by the creation of a linguistic representation of the world relatively autonomous with respect to experience. It has not been demonstrated, however, that this limitation pertains also to purely intellectual operations, whose purpose is to rationally process the perception stimuli with the use of mathematical models and to carry out cause-and-effect reasoning.

Bibliography

1. Altmann, Gabriel (2000) *Einführung in die quantitative Lexikologie.* Trier: Wissenschaftlicher Verlag Trier.

2. Buttler, Danuta (1967) "Koncepcje pola znaczeniowego." *Przegląd Humanistyczny* 2: 41-59.

3. Coseriu, Eugenio (1975) "Vers une typologie des champs lexicaux." *Cahiers de lexicologie* 27/2: 30-51.

4. Fillmore, Charles J. and Beryl T. S. Atkings (1992) "Towards a frame-based lexicon: the semantics of RISK and its neighbours." In Frames, fields, and contrasts, ed. Adrienne Lehrer, Eva Kittay, 75-102. Hillsdale, New York: Lawrence Erlbaum.

5. Fortescue, Michael (2009) *A Neural Network Model of Lexical Organisation.* London, New York: Continuum International Publishing Group Ltd.

6. Graszewicz, Marek and Dominik Lewiński (2007) "O nieistnieniu manipulacji." In Mechanizmy perswazji i manipulacji, ed. Grażyna Habrajska, 201-213. Łask: Oficyna Wydawnicza Leksem.

7. Hammerl, Rolf and Jadwiga Sambor (1990) *Statystyka dla językoznawców.* Warszawa: PWN.

8. Herdan, Gustav (1966) *The Advanced Theory of Language Choice and Chance.* Berlin *etc.*: Springer.

9. Jassem, Wiktor (1974) *Mowa a nauka o łączności.* Warszawa: PWN.

10. Johnson, Neal F. (1978) "Coding processes in memory." In Handbook of Learning and Cognitive Processes, vol. 6: Linguistic Functions in Cognitive Theory, ed. William K. Estes, 87-129. Hillsdale, New Jersey: Lawrence Erlbaum Associates Publisher.

11. Kay, Paul and Luisa Maffi (1999) "Color appearance and the emergence of evolution of basic color lexicons." *American Anthropologist* 101[4]: 743-760.

12. Köhler, Reinhard, Altmann, Gabriel and R. Piotrowski eds. (2005) *Quantitative Linguistik/ Quantitative Linguistics. Ein Internationales Handbuch / An International Handbook,* Berlin-New York: Walter de Gruyter.

13. Łobacz, Piotr and Matyja-Nawoja Mikołajczak (2002) *Skojarzenia słowne w psycholeksykologii i onomastyce psycholingwistycznej.* Poznań: Sorus.

14. Meyer-Eppler, Werner (1959) *Grundlagen und Anwendungen der Informationstheorie.* Berlin etc.: Springer.

15. Miller, George A. (1998) "Nouns in WordNet." In WordNet: An Electronic Lexical Database, ed. Christiane Fellbaum, 23-46. Cambridge, (MA): MIT Press.

16. Pawłowski, Adam (2003) "Struktura ilościowa pola leksykalnego kolorów." *Polonica* 22-23: 93-116.

17. Pawłowski, Adam (2006) "Quantitative linguistics in the study of colour terminology: A research report." In Progress in Colour Studies I: Language and Culture, eds. Carole P. Biggam, Christian C. Kay, 37-55. Amsterdam-Philadelphia: John Benjamins.

18. Pawłowski, Adam (2007) "Huffman coding trees and the quantitative structure of lexical fields." In Exact Methods in the Study of Language and Text. Dedicated to Professor Gabriel Altmann on the Occasion of His 75^{th} Birthday, eds. Peter Grzybek, Reinhard Köhler, 533-544. Berlin-New York: Mouton de Gruyter.

19. Piasecki, Maciej, Szpakowski, Stanisław and Bartosz Broda (2009) *A Wordnet from the Ground Up.* Wrocław: Oficyna Wydawnicza Polityki Wrocławskiej.

20. Pisarek, Walery (1967) *Das Wunder der Sprache: Probleme, Methoden und Ergebnisse der modernen Sprachwissenschaft.* Bern: Francke.

21. Rutkowska, Danuta, Piliński, Maciej and Leszek Rutkowski (1999) *Sieci neuronowe, algorytmy genetyczne i systemy rozmyte.* Warszawa: Wydawnictwo Naukowe PWN.

22. Sambor, Jadwiga and Rolf Hammerl eds. (1991) *Definitionsfolgen und Lexemnetze*. Lüdenscheid: Ram Verlag.

23. Tadeusiewicz, Ryszard (2000) "Wstęp do sieci neuronowych." In *Sieci neuronowe*, eds. Włodzisław Duch, Józef Korbicz, Leszek Rutkowski and Ryszard Tadeusiewicz, 3-28. Warszawa, Akademicka Oficyna Exit.

24. Tokarski, Ryszard (1984) *Struktura pola znaczeniowego*. Warszawa: PWN.

25. Tokarski, Ryszard (2006) "Pola znaczeniowe i ramy interpretacyjne — dwa spojrzenia na język." *LingVaria* 1: 35-46.

26. Trier, Jost (1973a) *Aufsätze und Vorträge zur Wortfeldtheorie*. The Hague-Paris: Mouton.

27. Trier, Jost (1973b[1931]) *Der Deutsche Wortschatz im Sinnbezirk des Verstandes*. Heidelberg: Winter.

28. Weisgerber, Leo (1950) *Vom Weltbild der deutschen Sprache*. Düsseldorf: Schwann.