Studies
in
Quantitative Linguistics
11

Issues
in
Quantitative Linguistics
2

edited by

Emmerich Kelih
Victor Levickij
Yuliya Matskulyak

RAM - Verlag

Issues in Quantitative Linguistics 2

edited by

Emmerich Kelih
Victor Levickij
Yuliya Matskulyak

Dedicated to Reinhard Köhler
on the occasion of his 60th birthday

2011
RAM Verlag

Studies in quantitative linguistics

Editors
Fengxiang Fan (fanfengxiang@yahoo.com)
Emmerich Kelih (emmerich.kelih@uni-graz.at)
Reinhard Köhler (koehler@uni-trier.de)
Ján Mačutek (jmacutek@yahoo.com)
Eric S. Wheeler (wheeler@ericwheeler.ca)

1. U. Strauss, F. Fan, G. Altmann, *Problems in quantitative linguistics 1*. 2008, VIII + 134 pp.
2. V. Altmann, G. Altmann, *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen.* 2008, IV+193 pp.
3. I.-I. Popescu, J. Mačutek, G. Altmann, *Aspects of word frequencies*. 2009, IV +198 pp.
4. R. Köhler, G. Altmann, *Problems in quantitative linguistics 2.* 2009, VII + 142 pp.
5. R. Köhler (ed.), *Issues in Quantitative Linguistics.* 2009, VI + 205 pp.
6. A. Tuzzi, I.-I. Popescu, G. Altmann, *Quantitative aspects of Italian texts*. 2010, IV+161 pp.
7. F. Fan, Y. Deng, *Quantitative linguistic computing with Perl.* 2010, VIII + 205 pp.
8. I.-I. Popescu et al., *Vectors and codes of text*. 2010, II + 161 pp.
9. F. Fan, *Data processing and management for quantitative linguistics with Foxpro*, 2010, V+233 pp.
10. I.-I. Popescu, R. Čech, G. Altmann, *The lambda-struture of texts.* 2011, II + 181 pp.
11. E. Kelih et al. (eds.), *Issues in Quantitative Linguistics Vol. 2*. 2011, IV + 188 pp.

Gedruckt mit Unterstützung der Karl-Franzens-Universität Graz

ISBN: 978-3-942303-07-1

RAM-Verlag
Stüttinghauser Ringstr. 44
D-58515 Lüdenscheid
RAM-Verlag@t-online.de
http://ram-verlag.de

Preface

This volume contains 14 contributions from international scholars from eight countries (Germany, India, Austria, Rumania, Slovakia, Czech Republic, Ukraine, USA). All contributions have their methodological background in quantitative linguistics, although with some overlap to other branches of linguistics like corpus linguistics, cognitive linguistics, and psycholinguistics.

In the first section, "*Lexical Semantics*", the contributions are devoted to the statistical analysis of collocations (substantives plus modal verbs) in German, to selected problems of the polysemy of semantic fields in English, to the diachronic analysis of semantic concepts in English, and to the statistical analysis of concept structures by means of associative analysis. The common background of these contributions is an attempt at analyzing concepts of cognitive linguistics and cross-cultural linguistics from a statistical and quantitative point of view. All in all, the contributions give a good overview of the methodological and theoretical state of the art in this particular branch of linguistics. This work is a starting point for the formulation of some general deductive hypotheses as usually required in quantitative linguistics.

The second section, "*Text and script analysis*", contains papers about the frequency structure of texts, based on the analysis of word form frequencies. On the one hand some text indicators are tested empirically, and on the other hand some new quantitative indicators of the thematic concentration in texts are presented. Syntactical issues are discussed on the level of sentence lengths (with a focus on data homogeneity) and selected problems of context comparison are analyzed by the means of combinatorial methods. Two papers are devoted to research of phonetics/semantics (sound symbolism) and to the quantitative analysis of an African script.

The third section, "*Words and word classes*", contains papers devoted to some morphological and morphosyntactical problems. One paper is a continuation of works within the Göttingen Project on Length Distribution in Linguistics and focuses on word length frequencies in Norwegian. Another paper discusses selected problems of a quantitative classification of part-of-speech systems. Furthermore, German substantive compounds are analyzed with regard to their frequency of occurrence and their combinations with some lexical classes. Finally, one further paper is devoted to the morphological and phonological word structure in Ukrainian and Indonesian, where an interpretation of the parameters of Menzerath's law is offered.

As a whole, the omnibus volume gives a good overview of the recent development in selected branches of quantitative linguistics and quantitative text analysis. We would like to thank our contributors for their effectiveness and productiveness in the process of editing this second omnibus volume of the "Studies in Quantitative Linguistics".

Emmerich Kelih

Contents

Words and word classes

Kelih, E. et al. (eds.) (2011), Issues in Quantitative Linguistics, Vol. 2, 1-21

Cognitive semantics and the corpus-based approach to concept structure

Nadia Yesypenko

Introduction

A nowadays undisputed fact about language is that it stands in very close relation to cognition. The problem lies in understanding the nature of that relation. In most varieties of Cognitive Linguistics it is an accepted fact that we can get some knowledge of cognition through the study of language, and vice versa. Edward Sapir wrote: "What is needed above all is a language that is as simple, as regular, as logical, as rich, and as creative as possible; a language which starts with a minimum of demands on the learning capacity of the normal individual and can do the maximum amount of work…" (Sapir 1949: 113).

The relation of language and cognition can be treated in different ways. If we interpret language as the influencing factor on cognition (language → cognition) we get something like the Sapir-Whorf hypothesis; the pair importance of language and cognition (language ↔ cognition) implies that a mutual influence exists. The dominant role of cognition (language ← cognition) sees language as a mere result of the operation of human general cognitive principles. There is another possibility to interpret the relation between language and cognition, that is as an equality (language = cognition). Language is partially autonomous, and the same holds for cognition: "Cognitive facts should not be understood as autonomous, any more than linguistic facts should" (Harder 1999: 198). At the same time a certain level of autonomy does exist. The arrow that points both ways seems to be the perfect option. This is undoubtedly an idea that indicates the direction of movement.

In keeping with such view, we admit the suggestion that at the lexical level, language and thought may mirror each other. On the linguistic side, naming of objects, events, or other entities is sensitive to influences such as cultural needs, contact with other languages, and sound changes, which drive meaning shifts including broadening, narrowing, differentiation, and reinterpretation of individual word meanings, and which can add words to or delete words from the language's lexicon (Hock and Joseph 1996). At the conceptual level, entities (objects, events, etc.) are represented as points in multidimensional feature space. They tend to form clusters in this space (Rosch and Mervis 1975). No fixed boundaries separate these clusters, and so conceptual categories are only implicitly defined. Further, the clusters formed may vary depending on feature weightings imposed by different contexts and task demands.

Conceptual groupings are not transmitted from generation to generation, but are formed to serve the demands of a particular task and are based on perception of features relevant to that situation. Of course, conceptual knowledge is culturally transmitted, but this knowledge is not about groupings per se; rather, it is about the nature of specific objects or types of objects.

The evidence for the reality of cultural norms and shared conceptions is provided by language and in particular by the meaning of the words. This principle asserts that language refers to concepts. In other words, the semantic structure of linguistic units (the meaning conventionally associated with words) can be equated with concepts. These conventional meanings associated with words are often treated as linguistic concepts or lexical concepts. The area of study known as cognitive semantics is concerned with investigating the relationship between experience, the conceptual system and the semantic structure encoded by language. Scholars working in cognitive semantics investigate knowledge representation (conceptual structure) and meaning construction (conceptualization). Cognitive semantics has employed language as the lens through which these cognitive phenomena can be investigated.

This idea is supported by Li and Gleitman in the following words: "Language has means for making reference to the objects, relations, properties and events that populate our everyday world. It is possible to suppose that these linguistic categories and structures are more or less straightforward mappings from a preexisting conceptual space, programmed into our biological nature. Humans invent words that label their concepts" (Li and Gleitman 2002: 266). The guiding principles of cognitive semantics advocate that words are treated as "points of access" to vast priorities of knowledge relating to a particular concept. Meaning construction is equated with conceptualization, a dynamic process whereby linguistic units serve as prompts for an array of conceptual operations (Evans 2006).

Linguistic semantics provides a rigorous methodology for decoding such meanings and consequently for elucidating the tacit assumptions which are linked with them. In my own reflection the following methods play the parts of the pivot-stones of the entire analysis of the conceptual mechanism and the concept formation in modern linguistic studies.

George Lakoff's reappraisal of metaphor

The metaphor has been seen within the Western scientific tradition as purely a linguistic construction. The essential thrust of Lakoff's work has been the argument that metaphors are primarily a conceptual construction, and indeed are central to the development of thought. He states that human's ordinary conceptual system, in terms of which we both think and act, is fundamentally metaphorical in nature (Lakoff 2003). Non-metaphorical thought is for Lakoff only possible when we talk about purely physical reality. For Lakoff, the greater the level of

abstraction the more layers of metaphor are required to express it. People do not notice these metaphors for various reasons. One reason is that some metaphors become 'dead' and we no longer recognize their origin. Another reason is that we just don't "see" what is "going on". For Lakoff, the development of thought has been the process of developing better metaphors. The application of one domain of knowledge to another domain of knowledge offers new perceptions and understandings. Metaphor plays a major role in structuring the way we think about the world.

The NSM (natural semantic metalanguage) method of semantic analysis

The key idea is a very simple one: in order to explain anything successfully, the terms in which we do the explaining must be simpler and more intelligible than the ones we are explaining. This means that the most effective technique for explaining the meaning of a word in a given context is reductive paraphrase, i.e. saying the same thing in simpler words. The reductive paraphrase procedure implies the existence of a terminal set of simple indefinable meanings. NSM researchers claim to have discovered this set of meanings (semantic primes) via a long and incremental program of experimentation with semantic analysis, extending from Wierzbicka (2007) through to Goddard (2007), and Gladkova (2010) continuing. From a theoretical point of view, the NSM metalanguage can be thought of as a highly disciplined and standardized subset of natural language: a small subset of word-meanings (63 in number), together with a subset of their associated grammatical properties. From a practical point of view, one can simply think of semantic primes, such as people, someone, do, say, want, know, think, good, bad, because, as a controlled vocabulary of plain simple words used for the purpose of semantic explication. NSM explications often consist of a combination of semantic primes along with other relatively simple, but non-primitive, words – termed ‘‘semantic molecules’’. Obviously, NSM English does not look exactly like what Sapir had in mind, because with its sixty-three lexical items, it is lexically poor rather than rich. It can indeed be regarded as a cognitive touchstone to all natural languages.

The methodology of cultural scripts

The theory of cultural scripts is the pragmatic ‘‘sister theory’’ of NSM semantics. Cultural scripts, by contrast, are stated in ordinary non-technical language, using the same metalanguage of semantic primes that is used in lexical semantic analysis. This does mean that cultural scripts are the same in nature as semantic explications. Cultural scripts are not paraphrases of actual lexical items, but ‘‘representations of cultural norms which are widely held in a given society and

are reflected in the language'' (Wierzbicka 2006: 56). ''They constitute a certain naıve axiology, that is, a naıve set of assumptions about what it is good and bad to do or say, and even to think and feel. Any given speech community has such shared assumptions, and although not everyone necessarily agrees with them, everyone is familiar with them because they are reflected in the language.'' One of the key concerns of work in the cultural scripts framework has been to de-naturalize the pragmatics of English. Despite the possible connotations of the word ''script'', it is important to stress that cultural scripts are not binding on individuals. They are not proposed as rules of behavior, or as descriptions of behavior, but as normative rules of interpretation and evaluation. It is up to individuals in concrete situations whether to follow (or appear to follow) culturally endorsed principles, and if so, to what extent; or whether to defy, manipulate, or subvert them, play creatively with them, etc. Whether or not cultural scripts are being followed in behavioral terms, however, the claim is that they constitute a kind of shared interpretive background.

The above-mentioned assumptions constitute theoretical part in cognitive linguistics studies. There is the danger of professional linguists failing in their responsibility to provide some practical advice and guidance that could help ease problems of concept stucture and verbalization.

Certainly the more we know about the structural possibilities of language, the more we shall spot points of verbal representation of mental units – concepts, and the more we shall be able to explain the nature of this process, but the first step – the recognition of concept distinctiveness – is intuitive. Can this process be made more objective? The present century has seen considerable progress in providing alternative methods for arriving at lexical decisions, aided by developments of statistics, computing and corpus linguistics.

While many recent cognitive-linguistic approaches to concepts verbalized in text were concerned in network-like categories representing the concept with many interrelated senses, corpus-lingistic approaches have remained rather agnostic as to how different word senses are related and have focused on distributional characteristics of different word senses representing a concept. This paper attempts to bridge the gap between these two approaches. Methods from these disciplines can be combined to improve search algorithms and demonstrate how cognitive linguistics can benefit from methodologies from corpus linguistics.

The existing methods of concept structure and verbalization analysis turned out to be too subjective and simplified, focusing more on semantic fields or synonymic lists of the words bearing the semantic meaning reflecting the concept sense. The actual cognitive processes underlying sense identification, distinction and its presentation in text result in subjective decisions. It is immediately obvious that early research on conceptual structure was largely based on a scholar's intuition and interpretation lacking a clear set of methodological decision principles. Therefore, corpus-based approach formulates strategies to provide a more objective foundation for resolving many issues by, for example, identifying

corpus-based traces of senses of the collocated or associated to the dominant lexical concept words that build up a concept structure.

Turning back to the problem of language-cognition relations, given that we draw conclusions on human cognition from the study of human language, it is interesting to investigate the embodiment of dominant cultural concepts of the English in the language of a literary text. George Lakoff (1987) witnesses the interrelation of language, cognition and culture. A detailed analysis of form, meaning and use of linguistic units leads to having to include cultural elements. The English language is culturally loaded. Wierzbicka (1996), while studying the linguistic expression of emotions, writes that most linguistic categories (words, constructions) referring to emotions in natural languages embody complex and culture-specific configurations of ideas about how thoughts, feelings may be related. Thus, the forms of concept expressions in English are much culturally co-determined. Though there is no full list of English cultural conceps, scientists often speak of nine dominant concepts in the English world view: home, freedom, privacy, fair play, restrain, gentleman, heredity, humour, common sense (Lechte 2003).

The present paper is concerned with the first concept on the list – the cultural concept HOME and its representation in literary text from the perspective of cognitive linguistics on the one hand and corpus-linguistics on the other hand. 36 English and American novels of the 18th to 20th centuries serve as the material of investigation

The goal of our study is to capture the concept HOME structure built up on the results of lexeme *home* concordance, coocurrence and collocations in the lexical context. We rely on Altmann's (2010) associative analysis based on co-occurrences of some exactly defined basic entities in a priori defined environments, V.V. Levickij's (2009) theory of collocations and adjoining words which amplify the lexical concept by bringing in their additional senses and emotional load to the concept sphere, J.R.Firth's (1957) study of the context-dependent nature of meaning with his notion of 'context of situation'. These approaches offer invaluable insight into lexicalized concept structure via the lexemes' usage and links within definite lexical environment. By exploring these focal points in depth we may be able to show the general organizing principles which lend structure and coherence to a cultural concept. These approaches are also based on computations which assure objective results of the study.

We put forward a hypothesis that extending our research to the lexical context we will notice that there is considerable diversity in what writers associate with the common and very general notions of "home".

We claim that in order to make sense of the world around us, a writer must be cognitively equipped with some image schemes of different world phenomena. The image schemes are mental constructs available to human cognition. We could further conclude that these schemes are employed for the conceptualization of such phenomenon as HOME. The conceptualization of this phen-

omenon is based on the author's experience associated with home and the realia of home determined by a given time period or culture. This being the case, we should also find some hints at that in language, and it can be assumed that also image schemes should leave traces in the verbalizations of the respective concept. In order to find such traces, one of the potential things to do is to consider all the words that adjoin the lexical concept. These words can have a straightforward connection to image schemes combined in HOME or they are closely linked with the phenomena related to HOME.

Once we speak of the adjoining words as the bearers of the concept's additional meaning, we can't but mention the adjoining words evolve associative meaning of the concept that will or will not correspond to the conceptual meaning of lexeme *home*. Conceptual meaning is the basic or universal meaning recognizable as a basic component of grammatical competence. It can be represented as a set of distinctive features described as semes. The operant features for "home" we find in The Living Webster Encyclopedic Dictionary of the English Language: "a dwelling or abode; a house or an apartment that is the fixed residence of a person, a family or a household; an accustomed or familiar neighbourhood; one's native city, region, country; the dwelling place of an animal; the habitat or seat; an institution for the care of the homeless, sick, infirm, orphaned; in various games the point which one tries to reach" [Dictionary 1982: 459]. Consider the following sentence:

Tom Jones had ridden one of Mr. Western's horses that morning in the chase; so that having no horse of his own in the squire's stable, he was obliged to go ***home*** *on foot: this he did so expeditiously that he ran upwards of three miles within the half-hour.(fielding)*

HOME ⟶ a habitat that is a residence of a person.

The semantic representation of the conceptual meaning is governed by the linguistic principles comparable to the paradigmatic and syntagmatic relations. The conceptual meaning is an inextricable essential part of the linguistic system. This property distinguishes conceptual meaning from its associative counterpart.

Associative meaning describes a composite of meanings drawn on certain mental connections expressed by language means. Such connections are based on the contiguities of real-world experience of the writer. Associative meaning will rely upon connotation and collocation. Connotation is the real-world value a writer associates with the concept. Real-world value is perceived in terms of socio-cultural principles, norms and rules. Connotation builds on the basic conceptual attributes various additional properties of the concept that we come to expect a referent to possess. Let us consider the expression "pleasant home" in the sentence:

Wemmick, I know you to be a man with a gentle heart. I have seen your ***pleasant home****, and your old father, and all the innocent cheerful playful ways with which you refresh your business life (Dickens).*

The expression "pleasant home" connotes one or several physical and emotional attributes such as: a nice stone house showing a social standing of the owner; the wealth and well-being of the family inhabiting the house. In the English culture a "house" is also equivalent to safety what is reflected in the saying: "My home is my castle".

Connotation is generally unstable. It varies considerably according to such factors as culture, historical period, social class and the general experience of a writer. Though in effect connotations are relatively peripheral meanings but they are a constituent part of the associative meaning of the concept.

Collocation is an umbrella term for the various instances of co-occurance of meaning. It refers to the sense a lexeme may acquire on account of the meanings of lexemes that tend to co-occur in similar environment (a sentence or a paragraph). This becomes apparent when we consider the following sentence:

*It is a most miserable thing **to feel ashamed of home**.*

A sense of "home" depends in this case on the verbal expression "to feel ashamed" – home is a place causing negative emotions, it is not a comfortable place to live in. If we proceed not only with the direct collocations but with the adjoing words used in the further sentences in the paragraph, we will see why "home" evokes unpleasant recollections in the author's mind:

*It is a most miserable thing to feel ashamed of **home**. There may be black ingratitude in the thing, and the **punishment** may be retributive and well deserved; but, that it is a miserable thing, I can testify (Dickens).*

The frequent usage of the adjoing noun "punishment" can be reinterpreted to mean that in the 17^{th}-19^{th} centuries punishment was a common place activity exercised on children at their homes.

Cases at hand show that associative meaning of the concept is wider than the conceptual one. Conceptual meaning is stable and invariable since it can be represented by means of a finite set of symbols, by their semantic features or semantic rules. In contrast, associative meaning is variable since it owes its validity to social and cultural factors. Associative meaning largerly depends on the real-world experience.

Though associative meaning refers to the ideal mental (cognitive) realm the methods of research appropriate for lexical concept realization should not rest on the researcher's intuitive conclusions. They require a corpus-based approach that allows quantitative methods to measure a degree of association between a concept and adjoing words revealing a whole spectrum of the meanings hidden in the concept HOME.

Concept structure

We now know broadly what a *concept is*, but the descriptive task of the concept structure remains. We can only image the size of the task facing those who wish

to get to grips with the lexical structure of the concept.

When we talk about the structure of the concept we are referring to the network of meaning relationships which bind lexemes, representing a concept together – what is know as its semantic structure. No lexeme exists in splendid isolation. As soon as we think of one lexeme (home), a series of others come to mind. Some of these lexemes help to identify home (pleasant, own, native), others relate to it closely in meaning (house, apartment, castle, cave, fortification, abode), others have a looser semantic connection (family, country, neighborhood), and there may be figurative or literary uses (Uncle Tom's Cabin), as well as a few individually author's or idiosyncratic associations (island, comfort, ship, childhood). If we probe all aspects of the semantic network which surrounds *home*, we build up a large number of connections. But if we look at a dictionary entry for *home*, we shall see very few of associations found in the novels represented there.

British linguist J.R. Firth drew attention to the crucial role of the lexical context which surrounds the lexeme, when we analyze its meaning. He said: „You shall know a word by the company it keeps" (Firth 1957: 11). This can be seen from the concordance. It shows the occurrence of *home* with enough context before and after to enable to see how this item is being used in each. The various meanings of the verbalized concept begin to appear when we examine the collocations or adjoining words. There are two useful ideas in the study of concept structure through lexical contexts: there is a central lexeme or node, surrounded by a fixed amount of lexemes in the sentences – the span within which the search of collocations take place. The span in our examples is quite large, allowing 10 or so words on either side of the node. We look at the lexemes which are immediately adjacent to a node and at those which fall withing three or four places on each side of it. For abstract concepts, we need to examine quite a wide span, and to look at many examples of use, in order for clear lexical patterns to emerge. Computational help is essential in each cases.

We have got a remarkable range of lexemes adjoining *home*, which were divided into theme clusters of nouns, verbs and adjectives. The amount of adjacent lexemes in the novels of the XVIIIth century is the smallest – 5302 units, in the novels of the XIXth century there are 9475 adjoining lexemes, in the XXth century books we find 9553 lexemes round the node. We believe that the relations between different senses of the concept HOME are motivated both culturally and frame-semantically, and also they are corpus-based as they rely on an exhaustive analysis of their concordance. The author's choice of words to represent concept HOME is in part determined by the frequency of co-occurrence of senses of *home* with particular lexical units united into several clusters. Table 1 shows clusters of nouns, verbs and adjectives adjoined to the lexical concept HOME

Table 1
The Adjoining Words Clusters

Part of Speech	Cluster	Cluster
noun	1. family/marriage: *family circle, father, mother, husband, uncle, son*	9. furniture: *mantelshelf, oven, table, chairs, clock, dressing-table*
	2. house: *dwelling, castle, house, mansion, kitchen*	10. neighbourhood: *back-yard, forest, coast, city, garden*
	3. kingdom/empire: *emperor, king, empire, subjects, court, queen*	11. time period: *morning, year, day, night, hour, tomorrow, week*
	4. feelings/emotions: *aggravation, trepidation, caution, confidence, love*	12. animals/plants that live in the house: *poodle, cat, spider*
	5. country: *country, land, England*	13. processes: *laughing, sleeping, cleaning, cooking, loss*
	6. distance: *mile, distance*	14. events: *death, meeting, party, ball*
	7. people who dwell in the house: *master, servant, bachelor, watchman*	15. music: *piano, music, song*
	8. food: *tea, milk, water, meal, supper*	16. money: *money*
verb	17. motion: *run, return, enter, go, reach, near, walk, come*	22. possession: *have, possess, own, obtain*
	18. existence: *be, stay, settle*	23. observation: *see, watch, glance, look, follow*
	19. destruction: *burn, blow off, burst, murder*	24. mental activity: *recall, think, refresh, know*
	20. feelings: *feel, experience, desire, relieve*	25. housekeeping activity: *bring, take, give, cook, toil*
	21. physical activity: *eat, devour, carry, provide, spare, draw*	
adjective	26. positive: *pleasant, elegant, mysterious, excellent*	29. color: *white, dark, brown, light, yellow*
	27. negative: *ashamed, uncomfortable, unacceptable, dirty*	30. size/shape: *huge, big, little, round, speckled-legged, gigantic*
	28. material: *wooden, steel, silver, gold, glass*	31. age: *old, new*

The clusters represent notions caused by the mental interpretation of the concept HOME. The concept brought to the author's mind different associations that reflect and pass on ways of living and ways of thinking characteristic of a given society and they provide priceless clues to understanding of English culture.

Associative analysis

Here we follow Altmann's approach to define exact probabilities of co-occurences. Associative analysis is based on co-occurrences of central lexeme *home* in selected lexical environment. We use lemmas whose environment is sentences in the text of a novel. The analysis tries to find overt or covert associations and computes the probability of lemmas to occur simultaneously in a definite text. Each novel is analyzed separately. The formula used for computation is as follows:

$$P(X \geq x) = \sum_{j=x}^{\min(M,n)} \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

where N is the number of sentences in the text, M is frequency of lexeme *home* in the text, n is occurence of the most frequent noun/verb/adjective, x is a number of sentences where *home* and the most frequent noun/verb/adjective occur simultaneously. The resulting number is the probability that *home* and a noun/verb/ adjective occurred together in the text. The smaller this number, the greater the association between them. Tables 2, 3, and 4 show lexemes, associated with *home* in the novels of British and American writers. In some novels we did not find significant probability of co-occurrence (Charles Brockden Brown "Wieland", Susanna Haswell Rowson "Charlotte Temple", Samuel Woodworth "The Champions of Freedom" – novels of the XVIIIth century; Nathaniel Hawthorne "The Scarlet Letter" – a novel of the XIXth century).

The words associated with the central lexeme *home* in the novels of the XVIIIth century reveal people's values, ideas, and attitudes to their homes. For the English people of that time, who were famous seamen and dauntless adventurers home was an owned adobe where they returned to see their families. The prevailing associated lexemes are verbs of motion: return, go, come, find, welcome, reach, attend; among nouns we find lexemes denoting family: daughter, son, father, master; and adjectives of positive characteristic: good, dear, happy.

Table 2
Words associated with HOME in the novels of the XVIIIth century

Novels		Associated word	Part of speech	Probability
American novels	Catherine Sedwick "The Travellers. A Tale"	leave	verb	0,027668
	George Tucker "A Voyage to Moon"	return	noun	0,033492
	Royall Tylor "Alegire Captive"	be	verb	0,018125
		love	verb	0,034061
		send	verb	0,034061
		day	noun	0,034061
		master	noun	0,034061
		own	adjective	0,034061
British novels	Ann Radcliffe "The Mysteries of Udolpho"	be	verb	5,43E-06
		attend	verb	0,036213
		know	verb	0,016771
		return	verb	3,38E-08
		come	verb	0,00623
		reach	verb	0,043739
		go	verb	0,043739
		welcome	verb	0,002255
		feel	verb	0,043739
		bring	verb	0,000277
		parent	noun	0,043739
		return	noun	0,003815
		country	noun	0,036213
		la vallee	noun	0,0118
		dear	adjective	0,043739
		anxious	adjective	0,036213
		happy	adjective	0,036213
		little	adjective	0,0118
		only	adjective	0,036213
		native	adjective	3,85E-05
	Daniel Defoe "Robinson Crusoe"	stay	verb	0,011673
		carry	verb	0,00104
		cut	verb	0,003746
		lay	verb	0,003746
		resolve	verb	0,039136
		go	verb	6,91E-09
		come	verb	0,011122

		kill	verb	0,042625
		bring	verb	2,43E-08
		take	verb	0,036021
		going	noun	0,036021
		thing	noun	0,006086
		cave	noun	0,043382
	Henry Fielding "Tom Johes"	be	verb	3,2E-10
		find	verb	0,043341
		send	verb	0,043892
		return	verb	1,63E-10
		go	verb	1,18E-06
		come	verb	0,003845
		father	noun	0,040025
		return	noun	3,96E-05
		house	noun	0,040025
		good	adjective	0,043892
		small	adjective	0,036295
		own	adjective	0,017968
		young	adjective	0,011854
	Jonathan Swift "Gulliver's Travels"	carry	verb	0,038576
		return	verb	0,000267
	Laurence Sterne "The Life and Opinion of Tristan Shandy"	return	verb	0,033492
		get	verb	0,010048
	Oliver Goldsmith "The Vicar of Wakefield"	be	verb	0,00583
		resolve	verb	0,034851
		come	verb	0,003341
		return	verb	0,003341
		go	verb	0,034851
		receive	verb	0,034851
		family	noun	0,005244
		daughter	noun	0,003341
		son	noun	0,003341
		pleasure	noun	0,003341
		mile	noun	0,003341

Table 3
Words associated with HOME in the novels of the XIXth century

Novels		Associated word	Part of speech	Probability
American novels	George Eliot "Silas Marner"	carry	verb	0,003661
		sit	verb	0,011515
		keep	verb	0,035779
		think	verb	0,038352
		go	verb	0,001149
		come	verb	0,016219
		day	noun	0,005908
		door	noun	0,035779
		homestead	noun	0,035779
	Henry James "Europeans"	go	verb	0,031929
		take	verb	0,039238
		fortune	noun	0,033802
	Harriet Beecher Stowe "Uncle Tom's Cabin"	tell	verb	0,035913
		be	verb	0,041358
		may	verb	0,035913
		come	verb	1,04E-05
		go	verb	3,45E-05
		take	verb	0,000742
		get	verb	0,011603
		sell	verb	0,035913
		die	verb	0,035913
		wife	noun	0,015799
		children	noun	0,042344
		hymn	noun	0,035913
		happy	adjective	0,043181
		free	adjective	0,035913
	Jack London "Martin Eden"	go	verb	0,006606
		return	verb	0,035522
		have	verb	0,035522
		way	noun	0,011347
		little	adjective	0,035522
	Maria Susanna Cummins "The Lamplighter"	be	verb	2,43E-29
		carry	verb	0,044459
		would	verb	0,044459
		come	verb	1,03E-05
		go	verb	4,43E-09
		leave	verb	0,003362
		return	verb	0,003484

		hasten	verb	0,012056
		walk	verb	0,012056
		bring	verb	3,84E-05
		take	verb	0,001555
		get	verb	0,044136
		look	verb	0,036601
		day	noun	0,003484
		evening	noun	0,006523
		father	noun	0,032832
		uncle	noun	0,036601
		anxiety	noun	0,036601
		spirits	noun	0,036601
		way	noun	0,036601
		absence	noun	0,044459
		pleasant	adjective	0,044459
		little	adjective	0,032832
		safe	adjective	0,044459
British novels	Ch. Bronte "Jane Eyre"	have	verb	0,012895
		be	verb	0,000205
		send	verb	0,036061
		come	verb	0,000378
		feel	verb	0,036061
		night	noun	0,043456
		Thornfield	noun	0,036061
		little	adjective	0,043456
		own	adjective	0,036061
		ask	verb	0,036471
	Charles Dickens "Great Expectations"	be	verb	2,8E-05
		keep	verb	0,036471
		let	verb	0,036471
		go	verb	2,12E-25
		come	verb	1,22E-08
		run	verb	0,043798
		walk	verb	0,043798
		get	verb	0,036471
		feel	verb	0,044218
		have	verb	0,000555
		take	verb	0,003377
		see	verb	0,006842
		look	verb	0,036471
		night	noun	0,039657
		father	noun	0,036471

		preparation	noun	0,036471
		dinner	noun	0,003908
		reason	noun	0,036471
		ashamed	adjective	0,01197
		own	adjective	0,036471
		last	adjective	0,036471
	Jane Austin "Pride and Prejudice"	talk	verb	0,011529
		be	verb	9,46E-08
		stay	verb	0,0358
		be to	verb	0,0358
		return	verb	0,0358
		have	verb	0,042971
		day	noun	0,003668
		staying	noun	0,0358
		way	noun	0,0358
		little	adjective	0,0358
	Mary Shelly "Frankenstein"	return	verb	0,002947
		bring	verb	0,01015
	Oscar Wilde "The Picture of Dorian Grey"	go	verb	0,005135
		farm	noun	0,034061
		home	adjective	0,034061
	W.Collins "The Woman in White"	be	verb	6,84E-18
		remain	verb	0,011655
		wait	verb	0,043328
		sit	verb	0,035992
		go	verb	0,000372
		get	verb	0,035992
		care	noun	0,035992

In fact, we find a more diverse association range of words in the novels of the XIXth century which may well serve as an introduction to a whole system of attitudes in the English and American culture, a glimpse of which we can obtain by contemplating some associated words like family, children, wife, father, uncle, homestead. These words prove that English culture encouraged family values. Home is still perceived as a place to stay, sit, care, be, remain, bring, have, and more frequently a place to return, go, reach, walk, come. In the novels we notice adjacent words of time as it became a new value due to a greater speed of life in this century. Adjectives bring positive associations to Home as well as the assessment of its size and possession.

Table 4
Words associated with HOME in the novels of the XXth century

Novels		Associated word	Part of speech	Probability
British novels	D.H.Lawrence "The Rainbow"	be	verb	0,022534
		stay	verb	0,002386
		kiss	verb	0,036543
		go	verb	7,95E-24
		come	verb	4,56E-05
		leave	verb	0,017195
		return	verb	0,003934
		stop	verb	0,036543
		become	verb	0,036543
		take	verb	0,044351
		night	noun	1,38E-05
		time	noun	0,003934
		supper	noun	0,036543
		way	noun	0,003934
		homestead	noun	0,012017
		feast	noun	0,036543
		own	adjective	0,003086
	Iris Murdoch "The Italian Girl"	be	verb	0,00021
		go	verb	1,24E-05
		come	verb	0,033492
	James Joyce "Ulysses"	be	verb	0,025053
		make	verb	0,03635
		come	verb	1,22E-09
		go	verb	9,99E-05
		bring	verb	0,0004
		take	verb	0,03635
		night	noun	0,002966
		family	noun	0,043994
		way	noun	0,01189
		house	noun	0,01189
		beauty	noun	0,01189
		happy	adjective	0,03635
		home	adjective	0,001248
	Muriel Spark "Momento Morie"	be	verb	2,92E-07
		go	verb	0,000339
		go	verb	0,000355
		get about	verb	0,035757
		nursing home	noun	1,6E-14

		health	noun	0,035757
		St. Aubrey's Home	noun	0,011501
		mental	adjective	0,011501
		poor	adjective	0,042892
	Somerset Maugham "The Moon and Sixpence"	make	verb	0,034176
		go	verb	0,000249
	Virginia Woolf "To the Lighthouse"	be	verb	7,72E-05
American novels	Gertrude Stein "Three Lives"	tell	verb	0,035779
		be	verb	8,6E-16
		live	verb	0,000356
		stay	verb	0,038352
		find	verb	0,003661
		keep	verb	0,035779
		sit	verb	0,035779
		help	verb	0,042932
		can	verb	0,038352
		go	verb	5,66E-12
		come	verb	2,87E-06
		brag	verb	0,035779
		papa	noun	0,035779
		thinking	noun	0,035779
		friend	noun	0,035779
		work	noun	0,035779
		married	adjective	0,001137
	Ernest Hemingway "For Whom the Bell Tolls"	come	verb	0,000653
		go	verb	0,032377
		take	verb	0,009353
	Nelle Harper Lee "To Kill the Mockingbird"	say	verb	0,024737
		be	verb	0,042652
		stay	verb	0,0365
		make	verb	0,040687
		meet	verb	0,043873
		lead	verb	0,0365
		would	verb	0,018676
		go	verb	7,32E-23
		come	verb	9,77E-07
		walk	verb	0,007671
		get	verb	0,003918
		turn	verb	0,011989

		cross	verb	0,0365
		get	verb	0,03661
		take	verb	0,043873
		time	noun	0,003918
		afternoon	noun	0,011989
		evening	noun	0,011989
		dinner	noun	0,007671
		way	noun	0,000453
		work	noun	0,0365
		car	noun	0,0365
		school	noun	0,043873
		christian	adjective	0,0365
		homemade	adjective	0,0365
	F.S. Fitzgerald "Tender is the Night"	go	verb	9,58E-10
		come	verb	0,011386
		drive	verb	0,035582
		get	verb	0,035582
		have	verb	0,011386
		take	verb	0,011386
		bring	verb	0,035582
	Sinclair Lewis "Main Street"	be	verb	1,38E-25
		sit	verb	0,036543
		keep	verb	0,036543
		can	verb	0,01279
		must	verb	0,044351
		go	verb	3,54E-11
		come	verb	2,17E-05
		drive	verb	0,012017
		get	verb	0,012017
		have	verb	0,000674
		bring	verb	0,012017
		get out	verb	0,012017
		see	verb	0,044351
		way	noun	0,000134
		good	adjective	0,017195
		own	adjective	0,003436
	Toni Morrison "The Bluest Eye"	be	verb	2,57E-05
		come	verb	0,00031
		go	verb	0,005441
		take	verb	0,035132
		home	adjective	0,003436

The words associated with *home* can lead us to a whole complex of cultural values and attitudes in the English society expressed inter alia in text of fiction. The changeability of culture is also reflected in the lexicon: concept HOME has remained more or less stable in the novels of the XVIII-XIXth centuries, but in the XXth century its meaning has expended (as reflected in the range of its accompanied words use) – in accord with established changes in the prevailing conceptions concerning home, home relations, life in general.

Concept HOME in English is like one loose end which we have managed to find in a tangled ball of wool: by pulling it, we may be able to unravel a whole tangled ball of attitudes, values and expectations, embodied not only in lexeme *home*, but mainly in the accompanied words, collocations, words associated with *home*.

Conclusion

Before we summarize the most important points of this paper and briefly talk about possible extensions, one caveat is necessary. Language and cognition are interrelated. Language encodes our thoughts reflecting different world phenomena by using symbols. The meaning associated with a linguistic symbol relates to a mental representation termed a concept. Concepts derive from percepts. The range of perceptual information deriving from the world is integrated into a mental image. A mental image is created according to our experience, culture, society, etc. The meanings encoded by linguistic symbols refer to a mental representation of reality as construed by the human mind.

Language provides prompts for the construction of a conceptualization. Therefore, to study language is to study patterns of conceptualisation. Our study shows how the meanings of the adjacent associated words reveal mental images of the concept HOME as intended by the writer.

The approach using corpus evidences, statistical methods of data analysis and cognitive interpretation allowed us to find (1) which adjoining words are most frequent and which characterization and cognitive motivation is therefore most relevant to the concept HOME, (2) that some senses of the concept reflected in the adjoining words are not listed in the dictionaries. This is astonishing since these senses are not fully predictable but they are historically and culturally motivated. We interpret this as evidence that concepts being the elements of national world view absorb and reflect notions important for the given nation in the given time period. We can also claim (3) that cognitive analyses benefit from corpus-based perspectives.

References

Altmann, G. (1996). The nature of linguistic units. *Journal of Quantitative Linguistics 3, 1-7.*

Evans, V., Gree, M. (2006). *Cognitive Linguistics. An Introduction*. Edinburgh: Edinburgh Univerity Press.

Firth, J.R. (1957). *Papers in Linguistics 1934-1951*. London: Oxford University Press.

Gladkova, A. (2010). *Russian Cultural Semantics: Emotions, values and attitudes*. Moscow: Languages of Slavic Culture.

Goddard, **Cl., Wierzbicka A.** (2007). *Semantic primes and cultural scripts in language learning and intercultural communication*. In: Sharifian**,** F., Palmer, G.B. (eds.), *Applied Cultural Linguistics 105–124. Amsterdam: John Benjamins.*

Gries, S.Th., Stefanowitsch, A. (2006). *Corpora in cognitive linguistics: corpus-based approaches to syntax and lexis.* Berlin-New York: Mouton de Gruyter.

Hanks, P. (2000). Do word meanings exist*? Computers and the Humanities 34, 205-215*.

Harder, P. (1999). Partial Autonomy. In: Th. Janssen and G. Redeker (eds.), *Ontology and Methodology in Cognitive Linguistics. Cognitive Linguistics: Foundations, Scope and Methodology: 195-222*. Berlin/New York: Mouton de Gruyter.

Hock H.H., Joseph B.D. (1996). *Language History, Langauge Change, and Language Relationship. An Introduction to Historical and Comparative Linguistics*. Berlin-New York: Mouton de Gruyter.

Jackendoff, R. (1983). *Semantics and Cognition*. Cambridge, MA: MIT Press.

Lakoff G., Johnson M. (2003). *Metaphors We Live By*. Chicago: University of Chicago Press.

Lakoff, G. (1987). *Women, Fire and Dangerous Things: What Categories Reveal About the Mind.* Chicago: University of Chicago Press.

Lechte, J. (2003). *Key Contemporary Concepts: From Abjection to Zeno's Paradox*. London, Thousand Oaks, New Delhi: SAGE Publications.

Levickij, V.V. (2009). Words-associations and additional meaning. In: E. Kelih, V. Levickij, G. Altmann (eds.), *Methods of text analyses: 165-181*. Chernivtsi: RUTA

Li, P., Gleitman, L. (2002). Turning the tables: language and spatial reasoning. *Cognition 83, 265–294.*

McEnery T., Wilson A. (2001). *Corpus Linguistics: An Introduction*. Edinburgh: Edinburgh Univeristy Press.

Pinker, S. (1994). *The Language Instinct*. Harmondsworth: Penguin.

Rosch, E., Mervis, C.B. (1975). Family Resemblances: Studies in the Internal Structure of Categories. *Cognitive Psychology **7**(4), 573–605*.

Sapir, E. (1949). *Selected Writings of Edward Sapir in Language, Culture and Personality*. Berkeley: University of California Press

The Living Webster Encyclopedic Dictionary of the English Language (1982). Chicago: The English-Language Institute of America.

Wierzbicka, A. (1996). *Semantics: Primes and Universals*. Oxford: Oxford University Press.

Wierzbicka, A. *(*2006). *English: Meaning and Culture*. New York: Oxford University Press.

Kelih, E. et al. (eds.), Issues in Quantitative Linguistics, Vol. 2, 22-38

Diachronic investigation of the semantic field of 'SIZE' in English

Olga Soloviova

Introduction

The possibilities of application of quantitative methods in the sphere of diachrony are not as numerous as in synchrony, but still these methods are often used in historical lexicology (Best 1983; Köhler 2005). Mostly, while investigating semantic development of words, linguists may apply intuitive methods. The aim of the present article is to demonstrate the use of quantitative methods in diachronic investigations.

Modern language investigations are marked with attention to specific items that are considered carriers of diverse linguistic and extralinguistic information categorized and conceptualized in a certain way, namely the concepts (Evans 2007: 42). Unlike cognitive linguistics in semasiology similar items are titled semantic fields – conceptual structures segmented by means of words that the human consciousness 'throws over as a net' on a holistic conceptual continuum (Levickij 2006: 20).

Thus, the study of language units within a comprehensive approach with regard to their paradigmatic and syntagmatic relations allows the exploration the history of the concept – realization and distribution of its expression by means of lexical units in different periods. Traditionally it is believed that the study of vocabulary in terms of the field theory is carried out in two ways: (1) onomastic (the concept) and (2) semasiological (the words), with the second approach being dominant, identifying the lexical-semantic field as a group of words of one language with very close semantic relations (Karaulov 1972: 57).

The aim of onomasiological approach is to establish various lexical means through which such notions are verbalized. In diachronic perspective onomasiological approach focuses mainly on the constant changes in the ways we express the concept, and thus on the detection of periodic schemes of nomination changes or, in our case, groups of concepts. It results in determining certain semantic language processes. The onomasiological approach also assists in elaboration of the typology of motives underlying semantic changes related to specific groups of language units. According to A. Blank (2003: 6-8) while carrying out semasiological analysis the focus can and must be placed on either individual cases of semantic change or word formation, or idioms, or borrowings, etc., while in onomasiological studies all the mentioned phenomena should be considered simultaneously.

It is well known that methodology of semantic field's investigation was elaborated by J. Trier. But the methodology was based on an intuitive analysis, while quantitative linguistics makes it possible to employ more objective and reliable statistical methods. Thus, the study of the semantic field "SIZE" in English must surely begin and be further proceeded with the study of the semantic continuum represented by adjectives as lexical units. In other words, the purpose of the article is to examine the evolution of the semantic field in the history of the English language carrying out the corresponding onomasiological analysis by means of quantitative methods.

The category of dimension as a set of ontological and axiological features of real and imaginary objects existing in a broad sense of the concept considers adjectives as its essential component. In addition, being an integral part of a broader category of space, dimension can easily come to denote temporalness and other categories. Such transitions are based on significative factors – metaphorical, metonymic and synesthetic transfers – which combine together various constituents to determine the formation of a versatile concept "SIZE" in human consciousness.

Method

Supporting the view of V. Levitsky and O. Lech (2008: 45), in the present paper we consider the semantic field as a set of constituents which are similar or related by their meanings. Thus, the semantic field "SIZE" in English can be represented by a set of meanings that were selected for each adjective:

(1) "size according to the number of components", e.g. *large crowd*;
(2) "size of temporal duration": *long time*;
(3) "size of sound";
(4) "size of natural phenomenon";
(5) "size of feelings and emotions";
(6) "size in terms of illumination";
(7) "size of abstract concepts";
(8) "physical size";
(9) "size of area occupied";
(10) "size according to several dimensions";
(11) "size according to the components size": e.g. *small salt*;
(12) "size of material wealth";
(13) "size in terms of social status";
(14) "size in terms of skills";
(15) "size in terms of smell": e.g. *great odor*;
(16) "size according to age";
(17) "size/positive evaluation";

(18) “size/negative evaluation”;
(19) “size in terms of physical efforts”;
(20) “size in terms of colour saturation”: e.g. *deep blue*.

All these segments are represented by the subclasses of nouns, with which examined adjectives collocate in texts. The overall number of examples for all the periods of the English language is around 18,000 collocations written out of texts (total number of words around 5,000,000).

Correlation analysis is used to establish the structure of the field and interrelation between separate meanings. This technique is applied in various research works for investigation of semantic and conceptual fields (Levickij 1966; Musurivs’ka 1993).

The diachronic research of the semantic field “SIZE” in English conducted in the paper consists of the following steps:

1) All syntagmata (word combinations or sentences) with adjectives denoting size were written out from texts belonging to four main periods of the English language;

2) All nouns that collocate with examined adjectives were divided into 20 subclasses. It is considered that each subclass represents a certain meaningful segment of the semantic structure of adjectives;

3) Frequency distribution of all subclasses in texts was entered into tables that underwent statistic analysis. Namely, the correlation coefficient *r* was calculated as:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

where x_i and y_i are the frequencies indices of the individual criteria compared, $\bar{x}$ and $\bar{y}$ are average values of these criteria. The computation was done with the help of computer software.

4) The results of correlation analysis were entered into Tables 2, 3, 4 and 5. Every time a certain number of constituents existing in a given period is entered into a table. To establish the position of each constituent within the structure of the investigated field one more "invariant" constituent was added. It includes the frequency of all other components and the correlations between this invariant and constituents determine the closeness to the center of the field (Levickij 2008: 47). All other calculated coefficients of correlation show the strength of ties between field constituents. The actual tie strength was considered ***strong*** within the range <0.68; 1> and as ***medium*** if an indicator was higher than 0.55 for Old English (OE); ***medium ties –*** between 0.47 and 0.59, ***strong ties –*** higher than 0.59 for Middle English (ME) and Early Modern English (EModE); and coefficients

higher than 0.58 indicate ***strong ties***, while coefficients between 0.46 and 0.56 ***medium ties*** in modern English (Levickij 2004: 82). These ranges were established based on the number of correlating pairs (13 for OE, 18 for ME and EmodE, and 19 for modern English). Medium ties, thus, are those with the significance level $\alpha = 0.05$; for strong ties the significant coefficients are established at the significance level $\alpha = 0.01$. The corresponding numbers are taken from the correlation coefficients table (Levickij 2004: 172). In Tables 2, 3, 4, and 5 ties within the significant range are highlighted (strong in bold type and medium in italics).

To determine the structure of the field "SIZE" in each period three important parameters were taken into consideration: the usage frequency of each constituent in texts, its proximity to the center of the field, and the quantity and quality of ties with other constituents (Levickij 2008: 47). These indicators were identified as fundamental for the core constituents of the field.

The Old English Period

In the Old English period adjectives denoting size with the seme 'large' had certain inventory and frequency advantage. Thus, we analyze nine adjectives with the seme 'large' and only four with the seme 'small'. The corresponding numbers of usage frequency are 734 and 132. The frequency distribution of denotation of each segment by different adjectives in the texts of OE is presented in Table 1.

Table 1
Frequency distribution of field constituents in OE texts

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**inv.**
micel	79	11	11	8	56	2	81	28	10	2	8	296
grēat	0	0	0	0	0	0	0	6	1	0	3	10
rūm	2	1	0	0	0	0	6	9	9	6	0	33
lytel	16	23	2	2	3	1	11	20	4	3	0	85
smœl	0	0	0	0	0	0	0	10	2	0	2	14
long	1	47	0	1	7	0	11	21	0	7	0	95
dēop	0	1	0	0	1	0	18	3	6	18	0	47
heāh	0	0	0	1	1	0	25	18	11	10	0	66
brād	2	0	0	1	0	0	0	11	12	7	0	33
wīd	3	29	0	1	0	0	5	6	32	14	0	90
sīd	5	1	0	1	0	0	10	10	18	19	0	64
nearu	0	4	0	0	0	0	4	7	1	2	0	18
eng	0	0	0	0	0	0	5	3	4	3	0	15
∑	108	117	13	15	68	3	176	152	110	91	13	866

The correlation coefficients are given in Table 2.

In OE paradigmatic relations within the semantic field divide this field into three groups: “size of temporal duration” (segment 2), the pair “size of area occupied” – “size according to several dimensions” (segments 9–10), and all other constituents that are united by strong ties (segments 1, 3, 4, 5, 6, 7, 8, 11). The correlation coefficients between constituents range from 0.76 to 0.99 indicating strong ties between the elements of the field.

In the zone of distant periphery there are three constituents that are not associated with the core or close periphery. Among them a medium tie (r = 0.61) is characteristic for semantic segments “size of area occupied” and “size according to several dimensions” with common seme “dimension”.

Table 2
Correlation coefficients of semantic field constituents in OE

	1	2	3	4	5	6	7	8	9	10	11	inv.
1	-	0.1	**0.99**	**0.98**	**0.97**	**0.96**	**0.93**	**0.7**	0.07	-0.2	**0.87**	**0.94**
2		-	0.09	0.19	0.15	0.17	0.07	0.44	0.1	0.05	-0.07	0.35
3			-	**0.97**	**0.98**	**0.96**	**0.93**	**0.69**	0.02	-0.3	**0.88**	**0.93**
4				-	**0.94**	**0.94**	**0.93**	**0.77**	0.15	-0.2	**0.82**	**0.97**
5					-	**0.89**	**0.95**	**0.7**	0.01	-0.2	**0.89**	**0.94**
6						-	**0.85**	**0.74**	-0.02	-0.3	**0.78**	**0.88**
7							-	**0.7**	0.1	-0.03	**0.79**	**0.93**
8								-	-0.1	-0.2	0.5	**0.77**
9									-	***0.61***	-0.1	0.23
10										-	-0.4	-0.01
11											-	**0.75**
invar.												-

The combination of size and time in ancient times was an essential characteristic of such a category as “space”, since “any complete description of the space by primitive or archaic consciousness involves the definition of “here – now”, not just “here” ” (Mify 1992: 340). Although space and time in that case form an indissoluble unity – chronotope, a constituent “size of temporal duration” takes a separate position within the OE “SIZE” field.

According to the criteria applied to structural elements, the semantic field reconstructed in the OE period has core (C), cloze (CP) and distant (DP) peripheries and can be represented as shown in Figure 1.

As is clearly shown in Figure 1 in the core of the OE semantic field there are three constituents: “size according to the number of components”, “size of abstract concepts”, “physical size”. It is important to remember that the element “physical size” is, from the genetic point of view, primary, while the intensity and abstraction are derivative and secondary. In addition, one of the main mani-

festations of size semantics in the OE period characterizes the concept of "number", indicating the inseparability of the "size", "number" and "intensity" concepts at the initial stages of the language development. This unanimity is especially verbalized by adjectives *lytel* and *micel*.

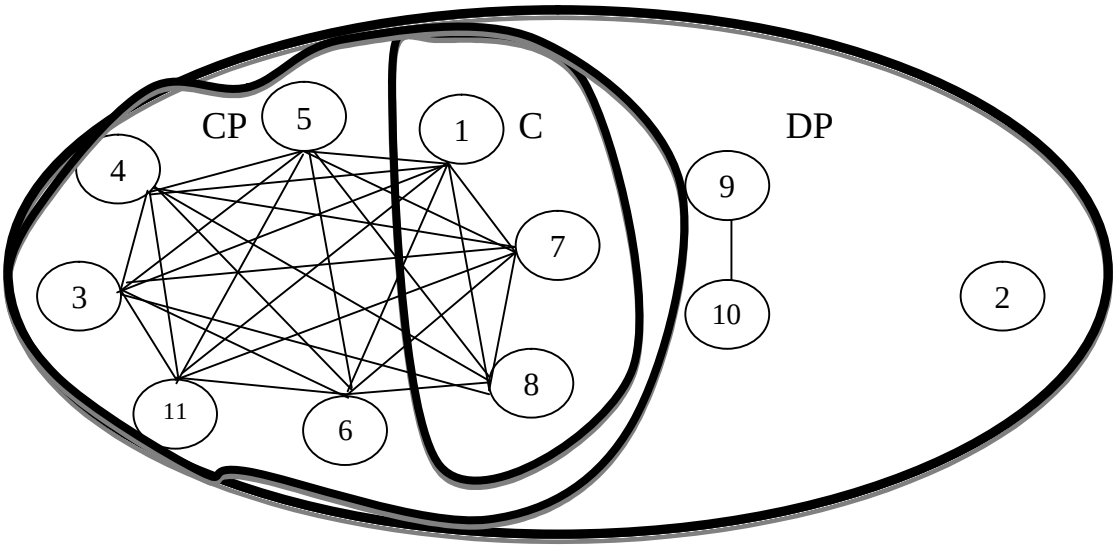


Figure 1. The structure of the Old English semantic field "SIZE"

The importance of the segments "size according to the number of components" and "size of abstract concepts" for the field in general is predetermined by an unclear division of quantitative and qualitative meanings of size. For instance, an adjective *micel* approaches the conveying the contemporary meaning of an adverb *much*: OE *teonan micelne* – modern English *much affliction*. Undoubtedly, there is an association of concepts "great" and "many", "many" and "very", "great" and "strong", "great" and "magnificent", "great" and "old", "great" and "thick". It means that the inclusion of components "great", "senior", "many", "strong", "important", and "significant" into the concept "SIZE" was not accidental, but a completely natural phenomenon (Levickij 1995: 118).

Syntagmatic relations within the field in the OE period have some specific features: 1) Different segments of the field are most closely related with the adjective *micel*; 2) dimensional field segments realize ties with adjectives which denote partial (*wīd*, *sīd*, *brād*) and general (*rūm*) size, which emphasizes the idea of similarity between "width" and "size"; 3) "size of temporal duration" was verbalized by adjectives *long*, *wīd*, *lytel*. Also the verbalization of different field constituents by various lexemes and partial differentiation of synonyms were characteristic for the period.

The Middle English Period

In the ME period the center of the field is transferred from the elements denoting physical and spatial size to the elements with the meaning of significance. In addition, changes in social and economic conditions led to enhancing the number of concepts that can be characterized in terms of size. Thus, in ME we differentiate between 16 constituents of the semantic field. The correlation between them is given in Table 3.

As we can see from the table the majority of significant ties are strong ($0.58 < \mathbf{r} < 0.99$). Medium ties are characteristic for segments 8 and 16 for which **r** ranges between 0.51 and 0.54. Several constituents possess only one tie each (2, 9, 10, 11).

Examining the origin of constituents of the field "SIZE" at the initial stages of English history, we can conclude about equality of metaphorical and metonymic meaning transfers in the domain of dimension. Thus, most of the adjectives that denote the quality of specific items, in our opinion, are used metaphorically to refer to abstract concepts. For example, the segments "size of feelings and emotions", "size in terms of skills", "size in terms of social status", "size of material value" developed metaphorically. Synesthetic metaphor prompts adjectives to acquire ability to verbalize "size of sound", "size in terms of illumination", and "size in terms of smell". The constituent "size according to the number of components" appeared as a result of metonymic transfer from the sphere of size to the sphere of quantity. In addition, the constituents that denote spatial size, amount of space occupied, size of substance components were established by means of metonymic transfers. The main condition for the development of the concept is the presence of stable context for the emergence of new meanings and needs of consciousness, which could serve as a potential prerequisite for the emergence of new lexical items and the development of additional meanings in existing ones.

The core group is comprised of constituents "size of abstract concepts" and "size of emotions and feelings" that led to the alienation of the physical size, along with parametric size and duration of time intervals.

Designation of quantity lost its priority within the field due to conversion of *moche* from the grammar class of adjectives into that of adverbs, though in some cases still preserving adjectival meaning.

Differentiation between "many" and "great" resulted in a gradual transition from a core constituent in the OE period to the distant periphery in contemporary English. Outside the periphery of the semantic field the following constituents occurred: 9 (size of occupied area), 11 (size of components) and 16 (size in terms of age). The reason for such changes in the structure of the field is the weakening of ties between its segments. Four elements (size of temporal duration, physical size, size of occupied area, size in terms of several dimensions) have the same

seme "distance" – temporal, spatial or physical – thus forming a separate model within the field.

Compared to the previous period, segment 11 ("size according to the components size") loses its significance removing from the group and keeping the tie (**r**=0.51) only on the level of segment "size in terms of age". The significance of this constituent during the previous period was provided by adjectives *smal* and *gret*, for which the meanings of "coarse-grained and fine-grained" were primary and which they lost as a result of generalization of meaning. Both adjectives in ME started to denote age in parallel structures: '*neighebores, bothe smale and grete*', '*Save al this compaignye, grete and smale!*' (Canterbury Tales). In general, the model can be represented as shown in Figure 2.

Paradigmatic relations within the field again show certain peculiar features of different segments. First of all, the segments possess strong ties with adjectives of both subgroups, but only either with a seme 'large' or with a seme 'small'. Thus, the first segment is associated with adjectives *gret*, *huge*; the tenth segment *large*, *wid*, *deep*; the third *smal*, *low*; and the sixteenth *litel*, *smal*. Thus, voice and age in this period were imagined as quiet and small, while dimensional size and number as *large*. Core field segments are characterized by strong ties with the dominant of lexical-semantic group *gret*.

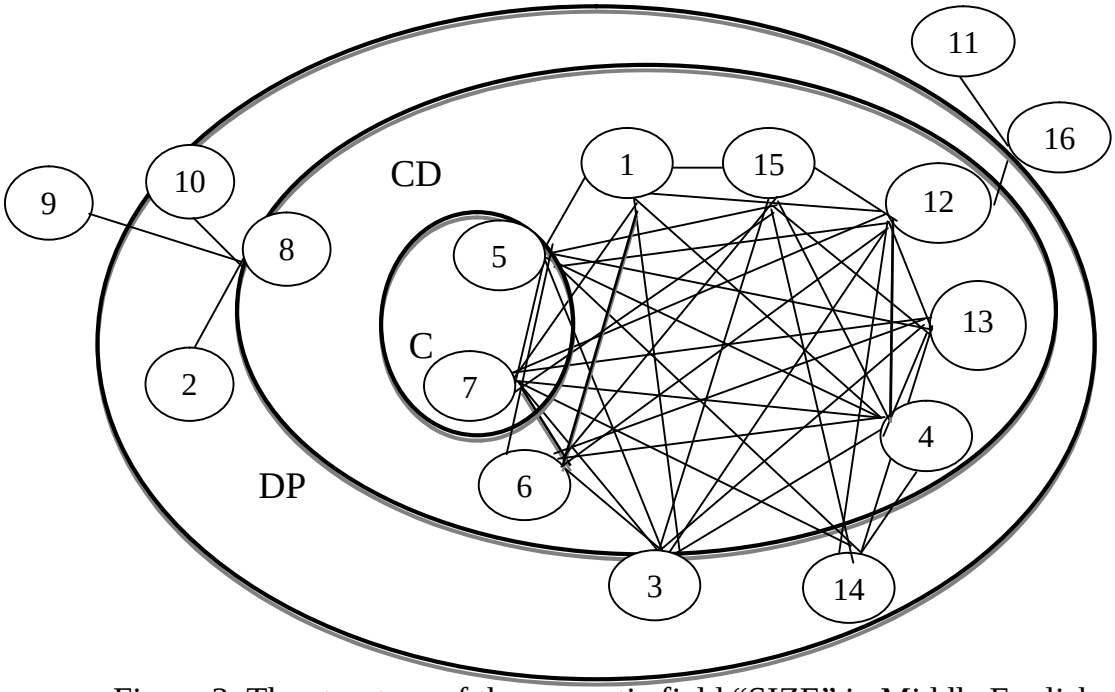


Figure 2. The structure of the semantic field "SIZE" in Middle English

Table 3: Coefficients of correlation between the segments of the ME semantic field 'SIZE'

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	inv.
1.	-	0.01	**0.66**	**0.89**	**0.96**	**0.85**	**0.96**	0.26	0.3	0.16	-0.03	**0.67**	0.49	0.49	**0.98**	0.2	**0.87**
2.		-	-0.1	-0.08	0.06	0.07	0.1	***0.52***	0.1	0.28	-0.1	0.1	-0.03	-0.03	0.02	0.19	0.25
3.			-	**0.68**	**0.65**	**0.77**	**0.66**	0.33	0.001	-0.02	0.45	**0.61**	**0.65**	0.37	**0.66**	0.34	0.59
4.				-	**0.84**	**0.66**	**0.84**	0.20	0.37	0.09	-0.14	**0.68**	**0.7**	***0.5***	**0.86**	0.01	**0.74**
5.					-	**0.83**	**0.99**	0.3	0.24	0.26	-0.07	**0.89**	**0.79**	**0.63**	**0.96**	0.25	**0.93**
6.						-	**0.85**	0.44	0.09	0.13	0.39	**0.76**	**0.58**	0.41	**0.85**	0.48	**0.75**
7.							-	0.35	0.22	0.28	-0.05	**0.87**	**0.81**	**0.66**	**0.97**	0.22	**0.91**
8.								-	***0.52***	***0.54***	0.42	0.48	0.34	0.37	0.24	0.49	0.47
9.									-	0.36	-0.16	0.25	0.1	0.1	0.21	0.07	0.32
10.										-	-0.11	0.37	0.28	0.33	0.18	0.21	0.32
11.											-	0.03	-0.1	-0.1	-0.1	***0.51***	-0.01
12.												-	**0.85**	**0.75**	**0.76**	***0.54***	**0.85**
13.													-	**0.91**	**0.70**	0.1	**0.75**
14.														-	***0.51***	0.06	0.62
15.															-	0.1	**0.83**
16.																-	0.3
inv.																	-

The Early Modern English Period

The analysis of the semantic field in this period is intended to show how the diversity of functioning and differentiating factors of dimensional adjectives tends to work within the current microsystem of that time weakening the relevant paradigmatic interconstituent ties compared with the previous period. The Early Modern English period lays the foundation for structure and functioning of the semantic field in Modern English: formation of new constituents, stabilization of semantic structures of lexical units, enhancing of differentiation degree between synonymous adjectives (e.g. *big* and *large*, *little* and *small*).

Apparently, peculiarities of social changes in society found their reflection in the structure of the semantic field, resulting in transfer of the segment "social status" to the core. The constituent "size of emotions and feelings", in contrast, moved from the core to the close periphery, reducing the frequency of use. In EModE the constituent "physical size" returned to the core of the semantic field.

As we did not find the examples of segment 15 realization in EModE texts we did not include it in our analysis. The highest number of significant ties for this period is 12 for segments (4, 5, 8, 9). The lowest number is two for segment 3. The general tendency that can be seen from the table is that the highest number of ties correlates with wider range of ties, while segments with a low number of ties usually enter into medium ties. Thus, for segment 3 **r** ranges from 0.58 to 0.67; for segment 4 **r** is from 0.52 to 0.94. The highest coefficients are characteristic for abstract segments: segments 7 and 5 (**r** = 0.94), segments 13 and 14 (**r** = 0.95), and segments 14 and 1 (**r** = 0.99).

The overall structure of the semantic field is presented in Figure 3.

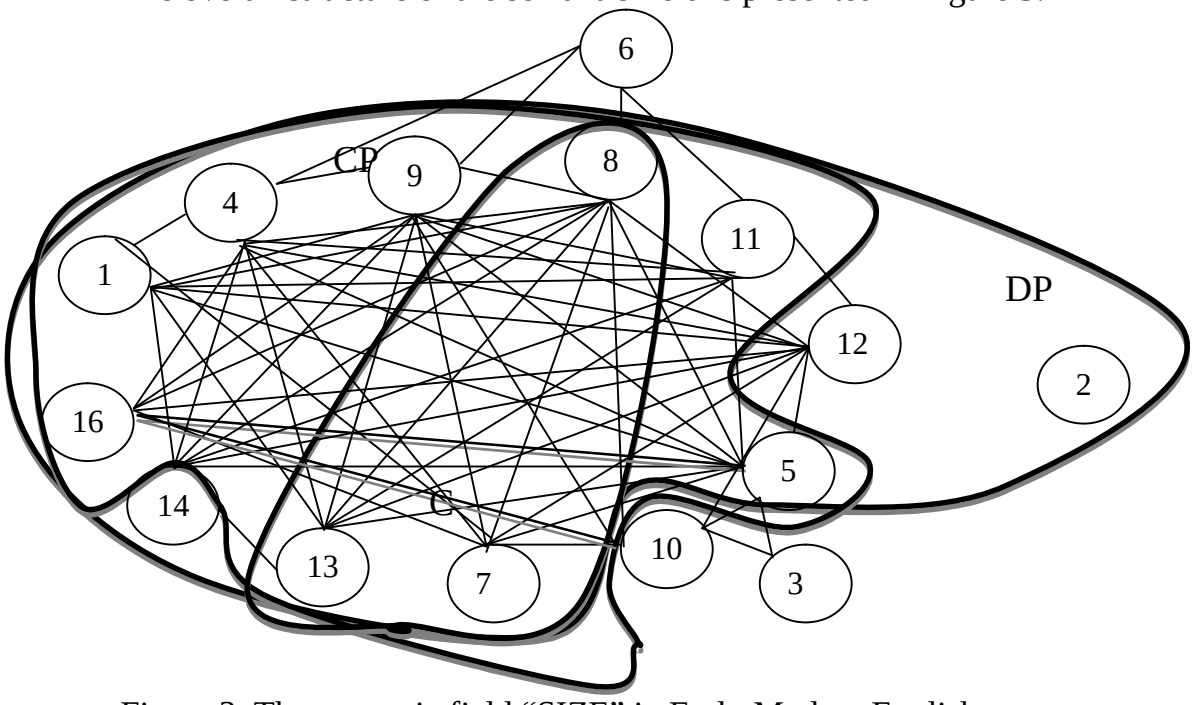


Figure 3. The semantic field "SIZE" in Early Modern English

The largest number of ties with adjectives is implemented by the constituent "physical size". All strong partners denote general and partial (namely horizontal dimension) size in antonymous pairs: *huge*, *large*, *big – small*; *long – short*; *brad*, *wide – narwe*.

In the sphere of distant periphery there are segments 2 and 14, while segments 3, 6, 10 and 15 turned out to be outside the peripheral zone. Apart from segments 2, 3 and 6 all field constituents were consolidated within the field possessing from 8 to 12 ties each. The triangular model "5-10-13" was possible due to the increase in significance of the adjectives denoting partial size in terms of vertical direction: *high/deep/low love – high/deep/low water – high/deep/low voice*.

Ties between field constituents and adjectives generally reflect the same trends as in the previous period: denotation of "size" of abstract concepts, feelings and emotions; collective concepts were mainly realized by the adjective *great*; verbalization of "size" in terms of age and separate components remained the function of adjectives with the seme "small"; segments "size in terms of social status" and "size in terms of skills" in EModE, in contrast, were characterized by strong ties with adjectives "big", "high", and "great".

In general, in the EModE period the following regularities can be observed: the share of metaphorical semantic field constituents increased including denoting "size" of abstract concepts, social status, abilities and skills, and vice versa, metonymic constituents were losing their positions, particularly in the subgroup of the general size, with the usage frequency of the segments "size according to several dimension" and "size of occupied area" significantly decreased.

The Modern English Period

Investigation of the semantic field "SIZE" in parallel with the group of lexical elements that verbalize the semantic continuum "SIZE" in modern English provides relevant information about how elements of the field interact inside and outside the examined field. The correlation analysis of segments is entered into Table 5.

The field core includes two segments that compete with each other during all periods; in modern English they are closest to the invariant – for the 8th segment (physical size) this distance is 0.46, for the 7th (size of abstract concepts) 0.5. It should be noted that in the modern period the distance between all constituents and the invariant is greater than in previous periods. Segments 3 and 11, 13, 14 and 20 appeared beyond the distant periphery. This phenomenon is caused by the decrease of lexemes that represent the semantic field segments in speech.

Table 4: Coefficients of correlation between the segments of the EModE semantic field 'SIZE'

	1.	**2.**	**3.**	**4.**	**5.**	**6.**	**7.**	**8.**	**9.**	**10.**	**11.**	**12.**	**13.**	**14.**	**15.**	**inv.**
1.	-	0.08	0.4	**0.82**	**0.9**	0.45	**0.91**	**0.74**	**0.85**	0.47	**0.77**	**0.82**	**0.92**	**0.99**	0.47	**0.93**
2.		-	0.1	0.08	0.06	-0.1	0.18	0.19	0.05	0.03	-0.01	0.13	0.01	0.06	0.2	0.21
3.			-	0.33	**0.67**	0.13	0.46	0.34	0.31	**0.58**	0.31	0.33	0.38	0.43	0.2	0.46
4.				-	**0.85**	***0.52***	**0.94**	**0.87**	**0.76**	**0.73**	**0.69**	**0.91**	**0.86**	**0.8**	**0.67**	**0.94**
5.					-	0.46	**0.94**	**0.78**	**0.8**	**0.67**	**0.74**	**0.82**	**0.85**	**0.89**	***0.53***	**0.94**
6.						-	0.49	**0.59**	***0.53***	0.19	**0.88**	0.39	0.44	0.44	0.23	0.5
7.							-	**0.84**	**0.80**	**0.66**	**0.74**	**0.87**	**0.92**	**0.91**	**0.58**	**0.99**
8.								-	**0.71**	**0.67**	**0.69**	**0.88**	**0.67**	**0.7**	**0.77**	**0.88**
9.									-	***0.54***	**0.79**	**0.77**	**0.75**	**0.83**	***0.5***	**0.83**
10.										-	0.27	**0.71**	0.47	0.42	**0.7**	0.67
11.											-	**0.6**	**0.72**	**0.78**	0.31	**0.76**
12.												-	**0.7**	**0.75**	**0.85**	**0.9**
13.													-	**0.95**	0.28	0.44
14.														-	0.36	**0.91**
15.															-	0.62
inv.																-

In contrast to the lexical-semantic group of adjectives whose elements over time tighten their ties and get closer to the core, the semantic field "SIZE" in English deconcentrates by reducing the number of core constituents (7, 8) and close periphery (4, 6, 9, 10, 12, 16) and increasing the number of elements outside the periphery. It can be corroborated by the distance of all segments from the invariant one (compare: in previous periods the minimum distance was 0.01, and in modern English 0.46). Compared to the ME period number of strong ties within the semantic field also decreased, the unity of the field model is lost and the largest number of ties (9) characterizes only one segment "intensity of natural phenomenon", with the remaining constituents possessing from 4 to 5 ties. The entire model of the semantic field is composed of discrete 3-, 4- or 5-component figures. All the models that comprise the semantic field can be represented by semes which unit the concepts: semes "bad/good" cause the ties between elements 15, 17, 18 (*r* ranges from 0.88 to 0.92); a seme "occupying some place" is common for field constituents 9, 8, 1 (*r* ranges from 0.52 to 0.78); a seme "great, outstanding" for 5, 7, 13 (*r* ranges from 0.52 to 0.87); a seme "diverse" for 1, 5, 7 (*r* ranges from 0.56 to 0.87); a seme "intense" for 5, 6, 7 (*r* ranges from 0.55 to 0.87). These potential semes are definitely very important; they appear in a variety of reinterpretations of the scrutinized concepts and, therefore, in an expansion of size concept towards the direction of additional semantic formation. The majority of ties between segments range from 0.6 to 0.7.

In modern English the total number of syntagmatic relations reduced dramatically, reflecting the unification of the group in terms of marking different segments of the field. The high correlation coefficient is characteristic for: *deep* with the segment "intensity of colour", *narrow* – "physical size", "size in terms of several dimensions"; *shallow* – "size in terms of several dimensions"; *low* – "size of temporal duration"; *high* – "size of origin and social status"; *broad* – "the size of illumination". Such links are caused by either a small number of adjectives that represent a certain segment ("colour intensity" – *deep*, *high*) or narrow combinability of adjectives.

A few words must be added concerning the field constituents which emerged by means of synesthetic conversion: "size of temporal duration" (3), "size of illumination" (6), "size in terms of smell" (15), "size in terms of colour saturation" (20). Adjectives that belong to the group of visual perception can easily convert to denoting concepts that belong to other sensitive areas as synesthesia is based on interdependence of impressions created in response to various human feeling (Stern 1968: 322-325). From the point of view of the semantic field "SIZE" in the first two periods the segments that relate to synesthesia not only belong to the core of the field or the close periphery, but also realize strong ties between themselves, forming a microsystem inside the field. Later the role of synesthesia in adjectives reduced due to many metaphorical meanings (12, 13, 14, 17, 18). Synesthetic constituents lost their ties with other segments and retreated to the distant periphery.

Table 5: Coefficients of correlation between the segments of the Modern English semantic field 'SIZE'

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**inv.**
1	-	0.05	0.1	0.4	**0.56**	0.4	**0.83**	**0.68**	***0.52***	0.12	0.15	**0.6**	0.31	0.1	0.14	0.5	0.35	**0.6**	0.09	-0.2	***0.5***
2		-	-0.01	-0.04	-0.01	0.4	0.08	0.3	0.15	**0.75**	-0.12	-0.06	-0.2	-0.1	0.03	0.1	0.18	0.08	0.5	-0.1	0.38
3			-	***0.52***	0.43	0.26	0.36	0.03	0.3	0.4	-0.1	0.2	0.02	0.42	***0.49***	0.29	0.35	0.37	0.43	0.42	0.2
4				-	0.38	0.26	**0.57**	**0.53**	**0.8**	0.33	0.1	**0.83**	0.3	**0.85**	**0.72**	***0.5***	0.38	**0.53**	0.14	-0.01	***0.45***
5					-	***0.55***	**0.87**	0.03	0.07	0.37	0.12	0.2	**0.57**	0.33	0.22	0.12	0.23	***0.45***	0.42	0.36	0.32
6						-	**0.62**	0.23	0.24	**0.69**	-0.01	0.01	0.07	0.05	***0.47***	**0.56**	**0.71**	**0.71**	***0.47***	0.1	0.43
7							-	0.38	0.37	0.36	0.16	0.4	***0.52***	0.35	0.33	0.44	0.44	**0.7**	0.27	0.01	***0.5***
8								-	**0.78**	0.26	-0.1	**0.75**	-0.1	0.1	0.32	**0.58**	0.24	0.43	0.2	-0.2	**0.54**
9									-	0.39	0.08	**0.84**	0.02	***0.51***	**0.58**	**0.56**	0.33	0.4	0.28	-0.03	**0.49**
10										-	-0.1	0.13	-0.04	0.26	***0.51***	0.3	0.43	0.35	**0.8**	0.38	**0.47**
11											-	0.17	**0.76**	0.09	0.02	0.08	0.08	0.14	-0.2	-0.1	0.05
12												-	0.3	**0.58**	0.4	0.38	0.07	0.3	0.09	-0.08	0.43
13													-	0.42	-0.02	-0.1	-0.1	0.15	-0.1	-0.1	0.12
14														-	**0.54**	0.06	0.1	0.2	-0.1	0.01	0.2
15															-	**0.74**	**0.75**	**0.7**	0.34	0.23	0.34
16																-	**0.88**	**0.9**	0.18	-0.1	0.42
17																	-	**0.92**	0.2	-0.1	0.34
18																		-	0.16	-0.1	**0.44**
19																			-	**0.78**	0.34
20																				-	-0.01

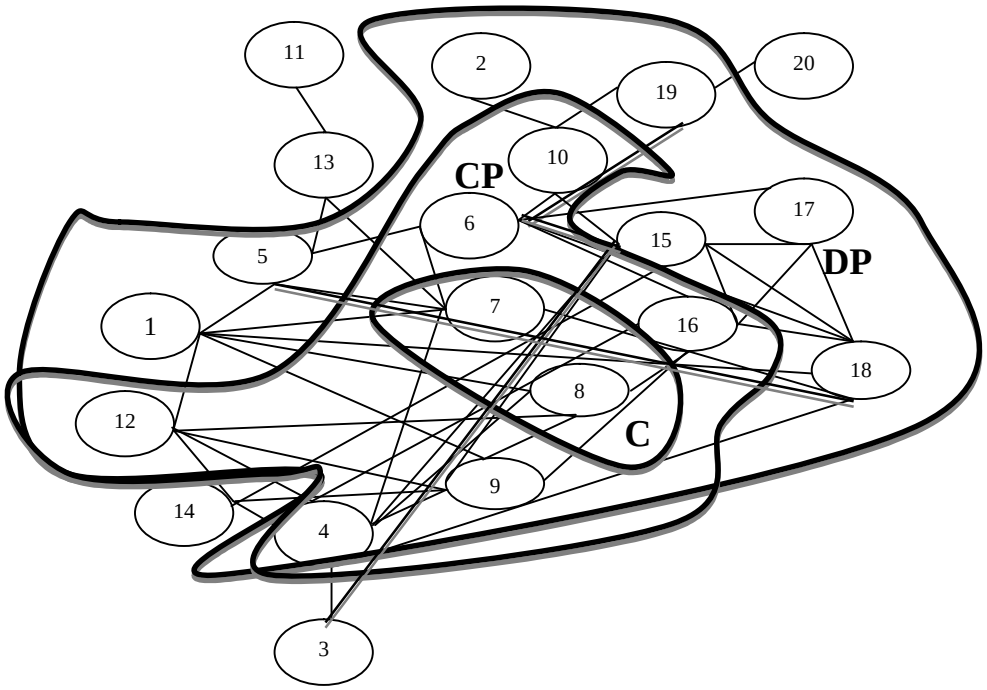


Figure 4. The representation of the semantic field “SIZE” in Modern English

Also interesting from the point of view of its evolution is the segment “size of temporal duration” (2) which does not possess more than one tie in any of the English language periods and always belongs to the distant periphery. This is due to the limited number of lexemes that can verbalize this semantic segment: syntagmatic relationship with the adjectives *little*, *long* and *short* were established only in modern periods.

The differentiation of meaning of each adjective that deepened in the course of the English language development led to considerable decrease in the number of paradigmatic and syntagmatic ties within the field. Thus, the whole structure of the semantic field “SIZE” is represented as a unit of separate models consisting of three to five units each of which can at the same time belong to several models. Some of these models as well as separate constituents can exist from OE times (like “size of temporal duration”), from ME (like size in terms of social position or skills) or appear in modern English (size in terms of colour saturation). But in modern English these models apparently become independent from the notion of “SIZE” and can be perceived as a different semantic unit or a part of another field (appraisal segment as a part of the semantic field “GOOD”).

Conclusions

The investigation of the semantic field “SIZE” as well as adjectival lexemes that

verbalize it was carried out with the help of statistical methods. The main tendencies observed in the course of the history of English are as follows. First of all, the unity of a field as a whole concept was lost and nowadays certain independent concepts can be perceived within the field; secondly, the study of lexemes shows that different field segments come to be attached to specific adjectives and lose their ties with other adjectives; thirdly, the segments themselves decrease the number of medium and strong paradigmatic ties due to differentiation of meaning of the adjectives.

The possibility to model the semantic field on the basis of its constituents' chronological depth and semantic derivation peculiarities is worth of further research. The semantic derivation of the field elements is mostly based on metaphor, metonymy and synesthetic metaphor. Different periods employ these transfers with different intensities and the appearance of new segments was based on extralingual as well as intralingual factors. The modern semantic notion "SIZE" appears to be the result of long and productive chain and radial derivation from the primary "physical size" to such modern components as "appraisal" or "colour saturation". For further investigation semasiological comparative analysis of different languages seems relevant. It may tell us if any notion development has deep extralinguistic and common foundations for different nations like synesthesia. Or it may prove that semantic evolution was provoked by certain paradigmatic or syntagmatic factors within one language.

References

Blank, A. (2003). Words and concepts in time: Towards diachronic cognitive onomasiology. In: R. Eckardt, K.v. Heusinger, Ch. Schwarze (eds.), *Words in time: diachronic semantics from different points of view: 37-66.* Berlin-New York: Mouton de Gruyter.

Best, K.-H., Kolhase J. (1983) *Exakte Sprachwandelforschung. Theoretische Beiträge, statistische Analysen und Arbeitsberichte.* Göttingen: Edition Herodot.

Evans, V. (2007). *A glossary of cognitive linguistics*. Edinburgh: Edinburgh University Press.

Karaulov, J.N. (1972). Struktura leksiko-semantičeskogo polja. *Filologičeskie nauki 1, 57-68.*

Köhler, R., Altmann, G., Piotrowski, R.G. (eds.) (2005). *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch – An International Handbook.* Berlin-New York: de Gruyter.

Levickij, V.V. (1966). *Istoriko-semasiologičeskoe issledovanie nekotorych grupp prilagateľnych v nemeckom jazyke (v sravnenii s anglijskim).* Moskva: Diss.

Levickij, V.V. (2004). *Kvantitativnye metody v lingvistike*. Černovcy: RUTA.

Levickij, V.V. (2006). *Semasiologija.* Vinnica: Nova Kniga.

Levickij, V.V. (2008). Konceptual'noe pole razmera v nemeckom jazyke i metody ego issledovanija. In: Levickij, V.V., Lech, O.S. (eds.), *Studia Germanica et Romanica: 40-57.* Doneck: DonNU.

Mify narodov mira: ėnciklopedija. Moskva: Boľšaja Rossijskaja ėnciklopedija.

Musurivs'ka, O.V. (1993). *Semantika prikmetnika v anglijskij movi.* Odessa: Diss.

Stern, G. (1968). *Meaning and Change of Meaning: with Special Reference to the English Language.* Bloomington: Indiana University Press.

Kelih, E. et al. (eds.), Issues in Quantitative Linguistics, Vol. 2, 39-56

Der Gebrauch von Substantiven mit Modalverben in gegenwärtigen deutschsprachigen Prosawerken

Viktor Drebet

In der heutigen linguistischen Forschung hat sich die Auffassung des Satzinhaltes als einer Ganzheit von propositionalem Gehalt/Aussagegehalt und pragmatischem Gehalt/Handlungsgehalt etabliert (vgl. Modell des Satzinhalts bei Polenz 1988: 93). Nach diesem Modell wird im propositionalen Gehalt etwas über Dinge und Erscheinungen der extralinguistischen Realität ausgesagt, die solcherart als Referenzobjekte auftreten. Der pragmatische Gehalt wird mit Begriffen wie Illokution und Perlokution angesprochen, wo Illokution und Perlokution als Sprecherhandlung (Illokution) mit einem bestimmten Wirkungsversuch (Perlokution) betrachtet werden können (vgl. das Modell ebenda). Der kommunikative Akt gehört somit zur Sprachkompetenz. Der Sprachkompetenz „soll die Leistung zugeschrieben werden, dass Produzent verbaler Äußerungen seine Gedanken und Absichten in sprachliche Strukturen umzusetzen vermag (Entkodierung)“ (Rüschel 1975: 14). Peter von Polenz macht diesbezüglich die Bemerkung, der Handlungsgehalt sprachlicher Äußerungen sei in der traditionellen Sprachwissenschaft meist vernachlässigt und in der Grammatik unter dem Stichwort „Modalität“ behandelt worden (vgl. Polenz 1988: 194). Er schreibt dabei, dass in der offiziellen Sprache Ichbezüge, Gefühls- und Einstellungsäußerungen, Interjektionen gemieden werden (vgl. ebenda). Pragmatische Sprachmittel, welche sich ihrerseits in der spontanen mündlichen Alltagssprache durch nonverbale/ nichtsprachliche Mittel wie Intonation, Gestik, Mimik, Körperbewegung ersetzen ließen, müsse man in vielen schreibsprachlichen Texten nach Polenz „mit der Lupe suchen“ (vgl. Polenz 1988: 194). Folglich sind die Sprecherhandlungen in vielen Fällen nicht explizit ausgedrückt, sondern sie sind nur implizit vorhanden, weswegen ohne Zweifel versucht werden muss, die verborgenen Intentionen im Text zu erschließen bzw. zwischen den Zeilen zu lesen. Polenz meint, die akademisierte Bildungs- und Öffentlichkeitssprache sei von einer Entpragmatisierung der Formulierungsweise gekennzeichnet, hier interessiere man sich mehr für Darstellung als für Appell und Ausdruck (vgl. Polenz 1988: 194). Unseres Erachtens müsste man nicht so entschieden mit der daraus zu ziehenden Folgerung über die Entpragmatisierung der Sprache vorgehen, auch wenn dies die akademisierte Bildungs- und Öffentlichkeitssprache betreffen mag. Sonst entsteht unwillkürlich der Eindruck, als unterschätze man die Modalität in ihrer funktionalen, dynamischen und zusammenhängenden Komplexität der syntaktischen und lexikalischen Einheiten, die sich auf die Bildung eines Inhalts mit bestimmten Intentionen des Textverfassers beziehen, weshalb der pragmatische Aspekt auch

bei der informativen Darstellung nicht verschwindet. Darüber hinaus weist Kozlovs'kyj mit Recht darauf hin, dass der strukturelle Satzinhalt eine subjektive Entstehungsgrundlage aufweist (Kozlovs'kyj 2008: 323). In diesem Zusammenhang halten wir es für zweckmäßig, die Definition der Modalität aus dem Metzler-Lexikon Sprache anzuführen, wonach die Modalität zum pragmatischen Gehalt gehört als eine „semantisch-pragmatische Kategorie, welche sich auf die Art und Weise der Stellungnahme des Sprechers zur Geltung des in einer Äußerung denotierten Sachverhaltes bezieht" (Metzler-Lexikon Sprache 1993: 395).

In der Rede eingesetzt integriert sich das Sprachzeichen in den jeweiligen Kontext je nach der bestimmten Situation, dem bestimmten logisch-gegenständlichen, denotativen Inhalt und korreliert mit konkreten grammatikalischen Formen und syntaktischen Konstruktionen (Jermolenko 1982: 12). Der Sprachforscher Hak charakterisiert die Situation als „Gesamtheit von Elementen, die im Bewusstsein des Sprechers in der objektiven Wirklichkeit im Moment des Sprechakts vorhanden sind und auf bestimmte Weise die Selektion der Sprachelemente bei der Bildung der Aussage an sich vorbestimmen" (Gak 1973: 358). Sich auf eine Situation orientierend unterstreicht der Forscher direkte Nominationen, d.h. semantische Strukturen, welche die Wirklichkeit isomorph wiedergeben, und abgeleitete Nominationen (Gak 1969: 80-83).

Unsere bisherigen Forschungen zeigen ihrerseits, dass im Falle der Auffassung des polysemen Wortes als eines paradigmatischen hierarchischen Gebildes mit einer Kette von Haupt- und Nebenbedeutungen, infolgedessen das polyseme Wort mit dem paradigmatisch monosemen Wort kontrastiert, das Substantiv in verschiedenen syntaktischen Konstruktionen in einer seiner Haupt- oder Nebenbedeutungen vorkommt; im Falle des monosemen Substantivs realisiert sich das monoseme Nomenpotential (siehe Drebet 2006; 2007; 2008; 2010).

Wir halten es für zweckmäßig, unsere Forschungen durch eine Eruierung möglicher Zusammenhänge zwischen substantivischen Realisationen und Modalverben zu erweitern, wo die letzten als Modalitätselemente für die Kenntlichmachung des Verhältnisses des Sprechers zur Aussage oder der Aussage zur Realität in Verbindung mit einem bestimmten Nominationstyp der Substantive im Satz fungierend auf die eine oder andere Weise auf die Erzielung bestimmter Verfassersintentionen gerichtet werden, sei es implizit oder explizit.

Dabei gehen wir von folgender Hypothese aus:

Unseres Erachtens müssten die meisten modalen Verhältnisse im Text explizit ausgedrückt werden, weil es vom Autor generell intendiert wird. Unsere Arbeitshypothese besagt, dass bei der Stellungnahme des Sprechers zur Geltung des modalen Sachverhaltes in syntaktischen Kontexten mit Modalverben diejenigen Substantive wohl selektiert werden müssten, mit welchen der komplexe Modalitätsausdruck nicht „mit der Lupe" gesucht wird. Das heißt, der komplexe modale Inhalt wird durch Substantive expliziert, die die Wirklichkeit meistens isomorph wiedergeben. Folglich müsste der pragmatische Gehalt bei der Bildung

der modalen Aussage die Selektion der Substantive mit kontextueller Realisation ihrer direkten Nomination und ihres monosemen Potentials vorbestimmen.

Im Kontext unserer Untersuchungen über die syntagmatischen Eigenschaften des Nomens im Deutschen (siehe Drebet 2006; 2007; 2008; 2010) setzen wir uns somit in der vorliegenden Arbeit zum Ziel, die Realisationen des Substantivs in seinen Hauptbedeutungen, Nebenbedeutungen/abgeleiteten Bedeutungen sowie die Realisationen des Substantivs als monosemes Sprachelements in den syntaktischen Rahmen zu studieren, in welchen die Modalität mit bestimmten Modalverben zum Ausdruck gebracht wird.

Für die Erfüllung dieses Ziels haben wir fünf deutschsprachige Prosawerke ausgewählt (Christa Wolf „Kindheitsmuster“, Heinrich Böll „Billard um halb zehn“, Hermann Hesse „Das Glasperlenspiel“, Martin Walser „Der Augenblick der Liebe“, Patrick Süskind „Das Parfum“). Nach dem Gesamtauswahl-Verfahren wurden jedem dieser Werke je 1000 Substantive entnommen, die sich in einem syntaktischen Kontext mit bestimmten Modalverben auf zehn von je zwanzig Seiten eines Werkes finden (die Gesamtzahl beträgt somit 5000 Substantive). Um ein möglichst objektives Bild vom bestehenden Zusammenhang zwischen Nomen-Realisation und Modalität zu gewinnen, haben wir es für sinnvoll gehalten, nur diejenigen Substantive der Analyse zu unterziehen, welche sich in den ausgewählten syntaktischen Rahmen realisieren, d.h. wenn die Substantive eine unmittelbare kontextuelle Integrierung in den syntaktischen Rahmen des Modalitätsausdrucks aufweisen. Da wir sowohl in der vorliegenden wie in unseren früheren Arbeiten (siehe Drebet 2006; 2007; 2008; 2010) von einer Hierarchie der sprachlichen Bedeutungen auf der paradigmatischen Ebene ausgehen, benutzten wir für die Feststellung der Monosemie, der direkten bzw. abgeleiteten Nominationen der kontextuell untersuchten Substantive das allgemein anerkannte erklärende Wörterbuch „DUDEN. Das große Wörterbuch der deutschen Sprache“ in zehn Bänden (DUDEN 1999), welches wir von hier an verkürzt DUDEN nennen. Die neben der Erklärung eines Wortes gegebenen Minikontexte helfen, den eventuellen Subjektivismus im Prozess der Unterscheidung zwischen direkten und abgeleiteten Nomen-Nominationen zu vermeiden, sie erleichtern es, die Art und Weise einer semantischen Transformation in der aktualisierten Rede zu bestimmen.

Da die statistischen Untersuchungsverfahren über viele Jahre hinweg ihre Prüfungen in verschiedenartigen wissenschaftlichen Forschungen im Ausland und in der Ukraine erfolgreich bestanden haben, sind wir uns der Notwendigkeit bewusst, nicht nur gewisse Zahlencharakteristiken eines Untersuchungsgegenstandes zu gewinnen, sondern auch mithilfe besonderer statistischer Verfahren die Zuverlässigkeit statistischer Resultate festzustellen (detaillierter über statistische Untersuchungsverfahren siehe Köhler/Altmann/Piotrowski 2005). Mithilfe des χ^2-Tests (Chi-Quadrat-Test) setzen wir uns zum Ziel, das Vorhandensein oder Fehlen der Zusammenhänge zwischen Realisationen eines bestimmten Nominationstyps der Substantive und Realisationen eines bestimmten Modalverbs

in den jeweiligen syntaktischen Konstruktionen zu ermitteln (detaillierter über Chi-Quadrat-Test siehe Zöfel 1992).

Somit ergab sich in der vorliegenden Arbeit für uns die Aufgabe, die Realisationsmöglichkeiten des substantivischen Wortes im syntaktischen Kontext mit den Modalverben zu untersuchen. Modalverben, wie bereits oben erwähnt, drücken die Modalität aus. Helbig und Buscha charakterisieren sie wie folgt: „Vorwiegend bedeutet diese Modalität eine Art, wie sich das Verhältnis zwischen dem Subjekt des Satzes und der im Infinitiv ausgedrückten Handlung gestaltet (Möglichkeit, Notwendigkeit, Erlaubnis, Verbot, Wunsch usw.). Daneben bedeutet sie jedoch auch eine Art, in welcher sich der Sprecher zu dem bezeichneten Vorgang verhält, vor allem seine Einschätzung der Realität dieses Vorgangs (Vermutung bzw. fremde Behauptung). Wenn die Modalverben die erste Funktion haben, spricht man von den Modalverben mit objektiver Modalität, wenn sie in der zweiten Funktion gebraucht werden, spricht man von der subjektiven Modalität der Modalverben“ (Helbig/Buscha 1996: 131). Zu den Modalverben der deutschen Sprache zählen die Autoren *dürfen, können, mögen, müssen, sollen, wollen* (Helbig/Buscha 1996: 131-137). Das Modalverb *lassen* charakterisieren sie als ein Verb mit Modalfaktor (Helbig/Buscha 1996: 187-188). Die postsowjetische Schule der deutschen Grammatik rechnet ihrerseits das Modalverb *lassen* schon zu den Modalverben (siehe Šendel´s 1988: 36; Birkengof/Romm/Uroeva 1980: 8-9; Tjagiľ 2002: 103-105). Wir werden dieses Modalverb auch zur Gruppe mit den Modalverben zählen, weil eine solche Einteilung unseres Erachtens eine Alternative darstellt, zwischen *lassen* als selbständigem Vollverb mit einem obligatorischen direkten Objekt und *lassen* als Modalverb mit dem Infinitiv I eines anderen Verbs im kontextuellen Gebrauch zu unterscheiden. Zum Beispiel:

1) *Ich lasse mein Gepäck auf dem Bahnhof* (Vollverb);
2) *Das Fenster lässt sich nur schwer öffnen* (Modalverb) (Tjagiľ 2002: 103, 105).

Folglich analysierten wir in den oben erwähnten Prosawerken die Realisation direkter, abgeleiteter Nomination der Substantive sowie die Realisation des monosemen Nomen-Potentials in denjenigen syntaktischen Rahmen, in denen objektive oder subjektive Modalität mit den Modalverben *dürfen, können, mögen, müssen, sollen, wollen, lassen* ausgedrückt wird.

Somit wurden unsere quantitativen Charakteristiken der statistischen Bearbeitung durch den oben erwähnten χ^2-Test unterzogen, um das Vorhandensein oder Fehlen des Zusammenhangs zwischen den substantivischen Realisationsmöglichkeiten und der mit Modalverben ausgedrückten Modalität zu ermitteln. Der minimale signifikante Chi-Quadrat-Wert bei der Anzahl der Freiheitsgrade $FG = 1$ und bei $P = 0{,}05$ beträgt in unserem Fall $\chi^2 = 3{,}84$.

Zum besseren Verständnis veranschaulichen wir unsere Berechnungen am Beispiel der Realisation der monosemen Substantive mit dem Modalverb *können*: 1) zuerst werden auf der Basis unserer Tabelle mit quantitativen Charakteristiken die Vierfeldertabellen für jeden Typ der substantivischen Realisation mit den Modalverben zusammengestellt. Für die Realisation der monosemen Substantive mit dem Modalverb *können* bekommen wir folgende Vierfeldertabelle:

Der Gebrauch der Substantive mit dem Modalverb *können* ist wie folgt darstellbar:

Modalverben	**Realisierte Nominationstypen der Substantive**				**Insgesamt**	
	monosem		**übrige**			
Können	504	**b**	1212	**a**	1716	**a + b**
Übrige	1088	**d**	2196	**c**	3284	**c + d**
Insgesamt	1592	**b + d**	3408	**a + c**	5000	**N**

wo *a, b, c, d* empirische Größen und *N* die Gesamtzahl aller Größen darstellen;

2) die in den Vierfeldertabellen gewonnenen Daten werden nach der χ^2-Formel bearbeitet:

$$\chi^2 = \frac{(ad - bc)^2 N}{(a + c)(b + d)(a + b)(c + d)};$$

$$\chi^2 = \frac{(1212 \times 1088 - 504 \times 2196)^2 \times 5000}{3408 \times 1592 \times 1716 \times 3284} = 7{,}34.$$

Somit beträgt das Ergebnis des χ^2-Tests 7,34, was dafür spricht, dass in der Gruppe der Realisationen monosemer Substantive mit dem Modalverb *können* die empirischen Größen die theoretisch erwarteten Größen statistisch signifikant überschreiten (bei der Anzahl der Freiheitsgrade $FG = 1$ und beim Wahrscheinlichkeitsgrad $P = 0{,}05$ beträgt der minimale signifikante Chi-Quadrat-Wert 3,84).

Die Ergebnisse der vorliegenden Untersuchung werden in Tabelle 1 gegeben, und zur besseren Veranschaulichung werden die quantitativen Charakteristiken dazu noch in Abbildung 1 dargestellt.

Tabelle 1
Zusammenhang zwischen den realisierten Nominationstypen der Substantive und der in syntaktischen Kontexten mit den Modalverben ausgedrückten Modalität

Modalverben in syntaktischen Kontexten zum Ausdruck der Modalität	**Realisierte Nominationstypen der Substantive**			**Insgesamt**
	Direkte Nomination	**Abgeleitete Nomination**	**Monosemes Substantiv**	
dürfen	68 $\chi^2 = 2{,}84$	36 $\chi^2 = 0{,}88$	28 $\chi^2 = 7{,}05$	132
können	792 $\chi^2 = 3{,}38$	420 $\chi^2 = 0{,}58$	504 $\chi^2 = 7{,}34$	1716
mögen	108 $\chi^2 = 5{,}76$	88 $\chi^2 = 7{,}59$	92 $\chi^2 = 0{,}0015$	288
müssen	368 $\chi^2 = 0{,}74$	204 $\chi^2 = 0{,}0018$	284 $\chi^2 = 0{,}85$	856
sollen	316 $\chi^2 = 0{,}46$	160 $\chi^2 = 1{,}86$	256 $\chi^2 = 3{,}88$	732
wollen	280 $\chi^2 = 6{,}43$	132 $\chi^2 = 0{,}13$	156 $\chi^2 = 5{,}65$	568
lassen	284 $\chi^2 = 5{,}92$	152 $\chi^2 = 2{,}55$	272 $\chi^2 = 16{,}45$	708
Insgesamt	**2216**	**1192**	**1592**	**5000**

Die sowohl in der Tabelle wie auch in der Figur dargestellten Ergebnisse sprechen dafür, dass die Gruppe der substantivischen Realisationen in den syntaktischen Rahmen mit dem Modalverb *können* zahlenmäßig die meist vertretene ist. Zur Veranschaulichung werden gleich unten Beispiele mit kontextueller Integrierung der Substantive in die gegebenen syntaktischen Rahmen angeführt:

1) „*Sie könne ihre mühsam erkämpfte Position in der Abteilung nicht durch eine solche Beziehung gefährden*“ (MW: 74).

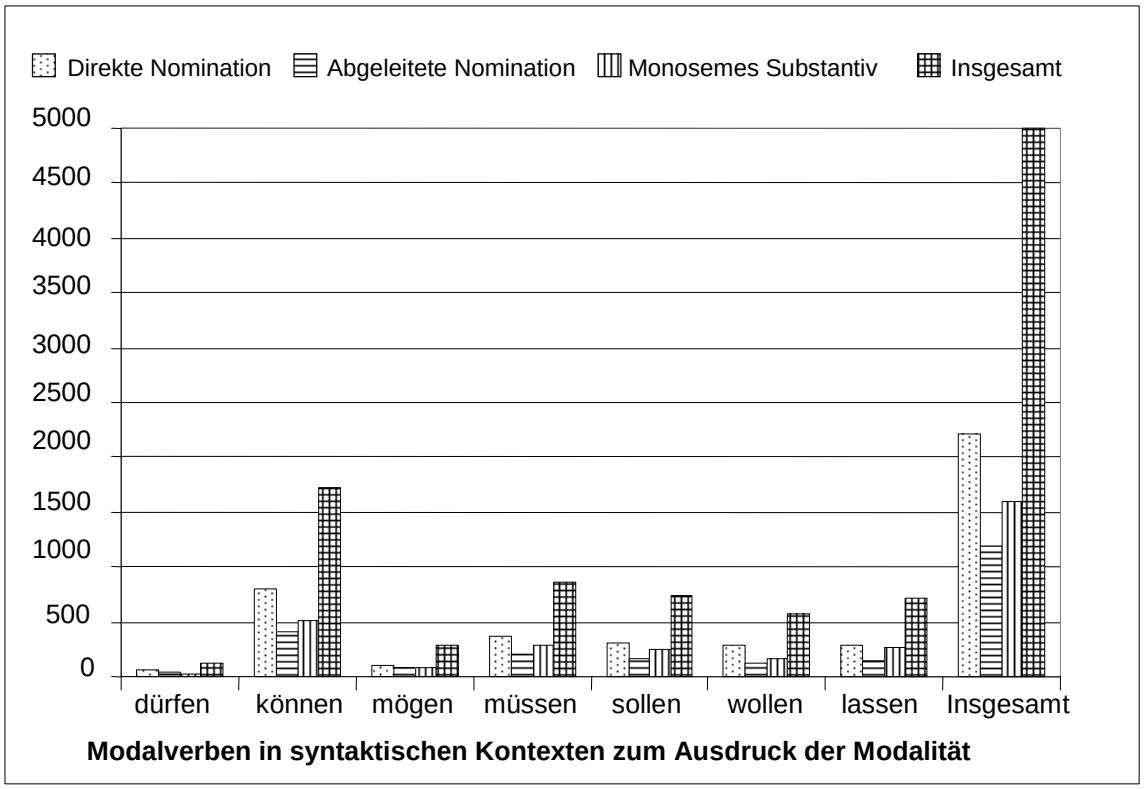


Abbildung 1. Graphische Darstellung des Zusammenhangs zwischen den realisierten Nominationstypen der Substantive und der in syntaktischen Kontexten mit den Modalverben ausgedrückten Modalität

In dem gegebenen Satz realisieren die Substantive *Position* und *Beziehung* ihre direkten Nominationen, d.h. ihre ersten lexikalischen Bedeutungen (die Position – 1.b) *bestimmte [wichtige] Stelle innerhalb einer Institution, eines Betriebes, eines Systems, einer vorgegebenen Ordnung o.Ä.;* die Beziehung - *Verbindung, Kontakt zwischen Einzelnen od. Gruppen*), und das Substantiv *Abteilung* realisiert seine abgeleitete Nomination, d.h. seine lexikalische Nebenbedeutung (die Abteilung – 2.c) *relativ selbstständiger Teil einer größeren Organisationseinheit (Unternehmen, Warenhaus, Krankenhaus u.a.*);

2) „*Hausarrest, Essensentzug, Strafarbeit, konnten sein Benehmen nicht ändern*“ (PS: 35).

In dem angeführten Beispiel sind die Substantive *Hausarrest, Essensentzug, Strafarbeit* monosem, folglich realisieren sie ihr monosemes lexikalisches Potential (der Hausarrest – *Strafe, bei der dem Bestraften verboten ist, das Haus zu verlassen*; der Essensentzug – *Entzug von Nahrungsmitteln (meist als Strafmaßnahme*; die Strafarbeit – *zusätzliche [Haus]arbeit, die einem Schüler, einer Schülerin zur Strafe aufgegeben wird*), und das Sub-

stantiv *Benehmen* integriert sich in den Kontext mit seiner direkten Nomination (das Benehmen – 1. *Art, wie sich jmd. benimmt; Verhalten, Betragen*).

Somit wurden in den syntaktischen Rahmen mit dem Modalverb *können* 1716 substantivische Realisationen festgestellt, davon gehören 792 Fälle zur Realisation direkter Nominationen, 504 Fälle gehören zur Realisation monosemer Substantive und 420 zur Realisation abgeleiteter Nominationen.

Ungeachtet der Tatsache, dass die gegebene Gruppe die zahlenstärkste ist, ergeben die statistischen Berechnungen mithilfe der χ^2-Formel den statistisch relevanten Zusammenhang zwischen der Realisation monosemer Substantive und der in syntaktischen Rahmen mit dem Modalverb *können* ausgedrückten Modalität, weil das ermittelte Ergebnis $\chi^2 = 7{,}34$ die minimal erwartete Summe überschreitet, welche in unserem Fall, wie oben erwähnt, 3,84 beträgt.

Wir stellen somit fest, dass sich nach den quantitativen Charakteristiken die höchste Zahl von Nomen-Realisationen in den syntaktischen Kontexten mit dem Modalverb *können* beobachten lässt, und zwar wenn die objektive Modalität mit dem Inhalt einer Möglichkeit und Fähigkeit bzw. wenn die subjektive Modalität mit dem Inhalt einer Ungewissheit ausgedrückt werden. Aber statistisch relevant sind in diesen inhaltlichen modalen Verhältnissen die Realisationen monosemer Substantive.

Als nächste hinsichtlich ihrer Zahlenstärke folgt die Gruppe der in syntaktischen Rahmen mit dem Modalverb *müssen* realisierten Substantive. Hier sind Beispiele dieser Realisationen:

1) „*Es musste eine Konferenz der Gesamtbehörde, unter Beiziehung auch der Altmeister, stattgefunden haben*“ (HH: 149).

In dem gegebenen Kontext wird das Substantiv *Konferenz* in seiner direkten Bedeutung realisiert (die Konferenz – 1. *Besprechung mehrerer Personen über fachliche, organisatorische o.Ä. Fragen*), und das Nomen-Kompositum *Gesamtbehörde* ist im DUDEN nicht vorhanden, d.h. dieses Wort ist eine Augenblicks-Zusammensetzung und gehört somit zu solchen Zusammensetzungen, „die als Wortbildungen des Textverfassers beim Formulieren zustandegekommen sind“ (Polenz 1988: 41). Sinnvoll ist unseres Erachtens eine semantische Analyse des Grundworts dieser Zusammensetzung (Substantiv *die Behörde*), welche uns in Folge mit vollem Recht ermöglicht, die ganze Zusammensetzung zu der Realisation des Substantivs mit einem monosemen Potential zu zählen (die Behörde – **a)** *staatliche, kommunale od. kirchliche Dienststelle, Verwaltungsorgan*). Das nächste Substantiv *Beiziehung* ist auch monosem (die Beiziehung – *das Beiziehen; zu Rate ziehen; in einem anstehenden Fall um sachverständige Äußerung, Behandlung od. klärende Bearbeitung bitten*), und das Substantiv *Altmeister* realisiert sich in seiner abgeleiteten Nomination (der Altmeister – 2. *bedeutendster, als Vorbild geltender Vertreter eines Berufszweigs od. Fachgebiets; Nestor, Senior*);

2) „*Man musste nicht alles und jedes gleich an die große Glocke hängen*“ (ChW: 62).

In diesem Satz haben wir es mit einer festen Verbindung von Wörtern zu tun, d.h. mit einer Redewendung *etwas an die große Glocke hängen*, welche bildlich oder metaphorisch gebraucht wird, aber auf der Ebene des Lexikon-Paradigmas wird diese Redewendung mit der ersten Bedeutung des entsprechenden Substantivs verwendet (die Glocke – 1.a) *aus Metall bestehender, in der Form einem umgedrehten Kelch ähnlicher, nach unten offener, hohler Gegenstand, der durch einen im Innern befestigten Klöppel zum Klingen gebracht wird*; **etw. an die große G. hängen** (ugs.; *etw. [Privates, Vertrauliches] überall erzählen*). Das heißt, wir beobachten auch direkte lexikalische Nominationen, welche feste metaphorische Redewendungen bilden können.

Somit wurden von uns in den syntaktischen Kontexten mit dem Modalverb *müssen* 856 substantivische Realisationen festgestellt, davon gehören 368 Fälle zur Realisation direkter Nominationen, 284 Fälle gehören zur Realisation monosemer Substantive und 204 zur Realisation abgeleiteter Nominationen. Wie die statistischen Berechnungen ergeben, hat kein realisierter Nominationstyp der Substantive die statistische Relevanz gezeigt. Das heißt also, dass in syntaktischen Kontexten mit dem Modalverb *müssen* zum Ausdruck objektiver Modalität mit dem Inhalt einer Notwendigkeit bzw. zum Ausdruck subjektiver Modalität mit dem Inhalt einer Gewissheit und Überzeugung kein statistisch relevanter Zusammenhang zwischen den substantivischen Realisationen und dem gegebenen Modalverb festgestellt wurde.

Die dritte Gruppe hinsichtlich ihrer quantitativen Charakteristiken ist die Gruppe der Nomen-Realisationen in syntaktischen Kontexten mit dem Modalverb *sollen*. Hier folgen Beispiele mit kontextueller Integrierung in die jeweiligen syntaktischen Rahmen:

1) *Unser Kastalien soll nicht bloß eine Auslese sein, es soll vor allem eine Hierarchie sein, ein Bau ...*“ (HH: 67).

Da *Kastalien* die Bezeichnung eines Ordens mit einer bestimmten dazugehörenden Kategorie von Menschen ist, realisiert sich das Substantiv *Auslese* in den jeweiligen kontextuellen Rahmen in seiner abgeleiteten Nomination (die Auslese – **2.** *die Besten aus einer Gruppe; Elite*), das Substantiv *Hierarchie* ist monosem (die Hierarchie - **a)** *[pyramidenförmige] Rangfolge, Rangordnung;* **b)** *Gesamtheit der in einer Rangfolge Stehenden*), und das Substantiv *Bau* aktualisiert seinerseits in dem gegebenen Kontext seine abgeleitete Nomination (der Bau - **2.a)** *Art, in der etw. gebaut, [kunstvoll] aus seinen Teilen zusammengefügt ist; Struktur*);

2) „*Aber er sollte seine Lehre bekommen, der präpotente Bursche!*“ (PS: 106).

In dem gegebenen Satz realisiert sich das Substantiv *Lehre* in seiner

abgeleiteten Nomination (die Lehre – **3.a)** *Erfahrung, aus der jmd. lernt, lernen kann*) und das Substantiv *Bursche* in seiner direkten Nomination (der Bursche - **1.c)** (abwertend) *männliche Person, der man Übles zutraut*).

Somit stellten wir in den syntaktischen Rahmen mit dem Modalverb *sollen* 732 substantivische Realisationen fest, davon gehören 316 Fälle zur Realisation direkter Nominationen, 256 Fälle gehören zur Realisation monosemer Substantive und 160 zur Realisation abgeleiteter Nominationen. Statistische Berechnungen ermittelten ein statistisch relevantes Ergebnis nur für die Realisationen monosemer Substantive (χ^2 = 3,88). Das heißt also, dass in den Wechselbeziehungen mit dem Modalverb *sollen* zum Ausdruck objektiver Modalität mit dem Inhalt eines Auftrags, einer Verpflichtung und Empfehlung oder zum Ausdruck subjektiver Modalität bei der Distanzierung des Sprechers von einer fremden Behauptung kontextuell nicht genannter Personengruppe über jemand oder etwas die Realisationen monosemer Substantive ihre statistische Relevanz aufweisen.
Sowohl aus der Tabelle wie auch aus der graphischen Darstellung lässt sich ersehen, dass hinsichtlich ihrer quantitativen Charakteristiken die Gruppe der Nomen-Realisationen in syntaktischen Rahmen mit dem Modalverb *lassen* als die vierte folgt. Hier sind Beispiele mit entsprechenden Realisationen:

1) „*Zwar, es gab Winter, da starben ihr von den zwei Dutzend kleinen Pensionären drei oder vier... So ließ sich mancher Ausfall verschmerzen*“ (PS: 27).

Die angeführten kontextuellen Rahmen zeigen, dass wir es in dem Satz, wo das Substantiv *Ausfall* gebraucht wird, mit der Realisation abgeleiteter Nomination des Wortes zu tun haben (der Ausfall - **2. b)** *Wegfall, Einbuße*);

2) „*... der violette Samt der Uniform ließ seine Gesichtshaut fast grün erscheinen*“ (HB: 82).

In den gegebenen syntaktischen Rahmen gehören alle Realisationen zu den monosemen Substantiven (der Samt - *feines Gewebe, meist aus Baumwolle, mit seidig-weicher, pelzartiger Oberfläche von kurzem Flor*; die Uniform - *bes. beim Militär u. bei der Polizei) im Dienst getragene, in Material, Form u. Farbe einheitlich gestaltete Kleidung*; die Gesichtshaut – *Haut des Gesichts*);

3) „*Die Staaten lassen ihre Leute also nicht unter allen Umständen töten, sondern nur unter genau beschriebenen Bedingungen...*“ (MW: 44).

In diesem Satz realisieren sich alle Substantive in ihren direkten Nominationen oder, mit anderen Worten, in ihren Hauptbedeutungen (der Staat – **1.a)** *Gesamtheit der Institutionen, deren Zusammenwirken das dauerhafte u. geordnete Zusammenleben der in einem bestimmten abgegrenzten Territorium lebenden Menschen gewährleisten soll*; die Leute – **1.** *mit anderen*

zusammen auftretende, als Menge o.Ä. gesehene Menschen; der Umstand – **1.** *zu einem Sachverhalt, einer Situation, zu bestimmten Verhältnissen, zu einem Geschehen beitragende od. dafür mehr od. weniger wichtige Einzelheit, einzelne Tatsache*; die Bedingung – **1. b)** *etw., was zur Verwirklichung von etw. anderem als Voraussetzung notwendig, gegeben, vorhanden sein muss*).

Somit wurden von uns 708 substantivische Realisationen in syntaktischen Kontexten mit dem Modalverb *lassen* festgestellt, davon gehören 284 Realisationen zu den direkten Nominationen, 272 Realisationen gehören monosemen Substantiven und 152 Realisationen kommen den abgeleiteten Nominationen zu. In dieser Gruppe wurden statistisch relevante Ergebnisse für die Realisationen monosemer Substantive (χ^2 = 16,45) sowie für die Realisationen der Substantive mit ihren direkten Nominationen bzw. in ihren Hauptbedeutungen (χ^2 = 5,92) ermittelt. Es sei darauf aufmerksam gemacht, dass der Chi-Quadrat-Wert für monoseme Substantive in der Gruppe mit dem Modalverb *lassen* der beste Wert von allen in der vorliegenden Arbeit untersuchten Größen ist. Im Allgemeinen stellen wir fest, dass die ermittelten statistischen Ergebnisse bezüglich ihrer Größenverteilung in der gegebenen Gruppe für die statistische Relevanz kontextueller Wechselbeziehungen bei den Realisationen des monosemen Nomen-Potentials sowie der direkten Bedeutungen des polysemen Nomen-Potentials mit dem Modalverb *lassen* zum komplexen Modalitätsausdruck mit dem Inhalt einer Veranlassung (Anordnung, Beauftragung), einer Erlaubnis, einer Zulassung sowie einer Möglichkeit/Unmöglichkeit, einer Tauglichkeit/Untauglichkeit sprechen.

Als fünfte Gruppe hinsichtlich ihrer Zahlencharakteristiken folgt die Gruppe der substantivischen Realisationen in syntaktischen Rahmen mit dem Modalverb *wollen*. Gleich unten werden entsprechende Beispiele angeführt:

1) „*Wer, bitte, will schon Ersatz sein*“ (MW: 74).
In diesem Satz realisiert sich das Substantiv *Ersatz* in seiner Hauptbedeutung (der Ersatz – **1.a)** *Person, Sache, die anstelle einer anderen Person od. Sache eingesetzt wird od. werden kann, deren Funktion übernimmt*);

2) „*Ich war kein Schuft, ... und wollte nicht mein Leben in Wohnzimmern verbringen...*“ (HB: 123).
Im gegebenen syntaktischen Rahmen treten die Substantive *Schuft* und *Wohnzimmer* als Realisationen monosemen Nomen-Potentials auf (der Schuft - (abwertend): *jmd., der gemein, niederträchtig ist; Schurke*; das Wohnzimmer – **a)** *Zimmer einer Wohnung für den Aufenthalt während des Tages*), und das Substantiv *Leben* zeigt die Realisation seiner Nebenbedeutung (das Leben – **2.b)** *Art zu leben, Lebensweise*).

Somit wurden von uns 568 substantivische Realisationen mit dem Modalverb *wollen* festgestellt, davon sind 280 Fälle die Realisationen direkter Nomi-

nationen, 132 Realisationen abgeleiteter Nominationen und 156 Realisationen monosemer Substantive. Als statistisch relevant zeigten sich Realisationen direkter Nominationen/Hauptbedeutungen der Substantive (χ^2 = 6,43) sowie Realisationen des monosemen substantivischen Potentials (χ^2 = 5,65). Die ermittelten statistischen Ergebnisse sprechen also für eine kontextuelle Relevanz der Realisationen sowohl monosemer Substantive als auch direkter Nominationen der polysemen Substantive mit dem Modalverb *wollen* zum komplexen Ausdruck einer objektiven Modalität mit dem Inhalt einer Absicht, eines festen Willens oder zum Ausdruck einer subjektiven Modalität bei der Distanzierung eines Sprechers von der Behauptung einer fremden Person bezüglich sich selbst.

Als sechste Gruppe hinsichtlich ihrer Zahlencharakteristiken folgt die Gruppe der substantivischen Realisationen mit dem Modalverb *mögen*. Zum Beispiel:

1) „*Die Karte mag vor der Einführung der norddeutschen Meile im Jahre 1868 gedruckt sein*“ (ChW: 373).

In diesem Satz realisiert sich das Substantiv *Karte* mit seiner abgeleiteten Nomination (die Karte – **6.** *kurz für Landkarte, Himmelskarte, Sternkarte*), die Substantive *Einführung, Meile, Jahr* realisieren sich ihrerseits mit ihren direkten Nominationen (die Einführung – **1.** *Das Einführen im Sinne als Neuerung bekannt machen u. verbreiten, in Gebrauch nehmen*; die Meile – **1.** *frühere Längeneinheit unterschiedlicher Größe (als Wegemaß)*; das Jahr – **1.** *Zeitraum von zwölf Monaten (in dem während 365 Tagen die Erde die Sonne einmal umläuft)*);

2) „*Auf die Vollendung des laufenden Spielkurses für Anfänger möge er ... keine Rücksicht nehmen*“ (HH: 148].

In diesem syntaktischen Rahmen zeigt das Substantiv *Vollendung* die Realisation seiner direkten Nomination/Hauptbedeutung (die Vollendung - **1.** *das Vollenden; das Vollendetsein*). Das nächste zusammengesetzte Wort *Spielkurs* ist im DUDEN nicht vorhanden, d.h. dieses Wort ist für den jeweiligen Kontext eine Augenblicks-Zusammensetzung, aber die semantische Analyse zeigt, dass das Grundwort *Kurs* unter Berücksichtigung des vorhandenen inneren Zusammenhangs mit dem Wort *Spiel* (*Spiel, das nach festgelegten Regeln durchgeführt wird*) den ganzen Inhalt dieser Zusammensetzung mit der abgeleiteten Nomination charakterisiert (der Kurs – **3.a)** *zusammengehörende Folge von Unterrichtsstunden o.Ä.; Lehrgang*). Der modale Inhalt dieses syntaktischen Rahmens wird dazu noch durch die Realisation des monosemen Substantivs *Anfänger* (der Anfänger - *jmd., der am Beginn einer Ausbildung od. einer Tätigkeit steht*) sowie die Realisation des Substantivs mit seiner direkten Nomination kontextuell ergänzt (die Rücksicht – **1.** *Verhalten, das die besonderen Gefühle, Interessen, Bedürfnisse, die besondere Situation anderer berücksichtigt, feinfühlig beachtet*).

Aus der Tabelle lässt sich ersehen, dass in syntaktischen Kontexten mit dem Modalverb *mögen* 288 substantivische Realisationen festgestellt wurden, davon gehören 108 Realisationen den direkten Nominationen, 88 den abgeleiteten Nominationen und 92 gehören den monosemen Substantiven. Statistisch relevante χ^2-Ergebnisse zeigten die Realisationen direkter (χ^2 = 5,76) und indirekter Nominationen (χ^2 = 7,59). Es sei darauf aufmerksam gemacht, dass von allen Realisationen abgeleiteter substantivischer Nominationen mit Modalverben einen statistisch relevanten Zusammenhang nur die mit dem Modalverb *mögen* zeigen.

Somit sprechen die statistischen Ergebnisse für eine kontextuelle Aktualität der realisierten direkten und abgeleiteten substantivischen Nominationen mit dem Modalverb *mögen* zum komplexen Ausdruck objektiver Modalität mit dem Inhalt eines höflichen Wunsches, einer Abneigung oder zum komplexen Ausdruck subjektiver Modalität mit dem Inhalt einer einräumenden Vermutung.

Als allerletzte siebte Gruppe hinsichtlich ihrer quantitativen Charakteristiken enthält unsere Klassifikation die Gruppe der substantivischen Realisationen mit dem Modalverb *dürfen*. Gleich unten folgen die Beispiele:

1) „*Die Natur war mit den Sinnen erfahrbar, studierbar, prüfbar. Und soweit sie nicht erkennbar war, durfte sie nicht in den Dienst der Erkenntnis genommen werden*“ (MW: 115).

Die gegebenen syntaktischen Rahmen zeigen, dass im Satz, in dem das Modalverb *dürfen* gebraucht wird, das Substantiv *Dienst* sich mit seiner direkten Nomination realisieren lässt (der Dienst – **1.b)** *Arbeitsverhältnis, Stellung, Amt*) und das Substantiv *Erkenntnis* mit seiner abgeleiteten Nomination (die Erkenntnis – **2.** *Fähigkeit des Erkennens, des Erfassens der Außenwelt*);

2) „*... und der Name der Gottheit durfte nicht genannt, nur gedacht werden*“ (HB: 63).

Im angeführten syntaktischen Rahmen realisieren sich beide Substantive mit ihren abgeleiteten Nominationen oder in ihren Nebenbedeutungen (der Name – **2.a)** *kennzeichnende Benennung eines Einzelwesens, Ortes od. Dinges, durch die es von anderen seiner Art unterschieden wird; Eigenname*; die Gottheit – **2.** *nicht eindeutig bezeichneter Gott bzw. Göttin*).

Somit wurden von uns in syntaktischen Kontexten mit dem Modalverb *dürfen* 132 substantivische Realisationen festgestellt, davon gehören 68 Realisationen zu direkten Nominationen, 36 Realisationen gehören zu abgeleiteten Nominationen und 28 Realisationen zu monosemen Substantiven. Die statistischen Berechnungen nach der χ^2-Formel ermitteln einen statistisch relevanten Zusammenhang der Realisationen monosemer Substantive in syntaktischen Kontexten mit dem Modalverb *dürfen* (χ^2 = 7,05).

Statistisch nachgewiesen wurde demnach in der zahlenmäßig kleinsten Gruppe unserer Klassifikation die kontextuelle Aktualität der Realisationen monosemer Substantive mit dem Modalverb *dürfen* zum komplexen Ausdruck objektiver Modalität mit dem Inhalt einer Erlaubnis, eines Verbots oder zum komplexen Ausdruck subjektiver Modalität mit dem Inhalt einer Wahrscheinlichkeit.

Ausgehend von den Ergebnissen unserer Arbeit lässt sich somit folgendes zusammenfassen:

1) in fünf von uns untersuchten deutschsprachigen Prosawerken gehört die größte Zahl substantivischer Realisationen zu den direkten substantivischen Nominationen oder Hauptbedeutungen der Substantive, den zweiten Platz hinsichtlich ihrer quantitativen Charakteristiken belegen die Realisationen monosemer Substantive, und den dritten, zahlenmäßig letzten Platz belegen Realisationen der abgeleiteten Nominationen oder Nebenbedeutungen der Substantive;
2) bei der Untersuchung von Realisationszusammenhängen zwischen Substantiven und Modalverben zum Modalitätsausdruck des Inhalts in den deutschsprachigen Prosawerken wurde festgestellt, dass sich die größte Zahl substantivischer Realisationen in syntaktischen Kontexten mit dem Modalverb *können* beobachten lässt, den zweiten Platz hinsichtlich ihrer Zahlenstärke belegen substantivische Realisationen mit dem Modalverb *müssen*, den dritten mit dem Modalverb *sollen*, den vierten mit dem Modalverb *lassen*, den fünften mit *wollen*, den sechsten mit *mögen*, und den siebten, zahlenmäßig letzten Platz belegen substantivische Realisationen mit dem Modalverb *dürfen*.
3) statistische Berechnungen mithilfe vom χ^2-Test für eine Ermittlung des Vorhandenseins oder Fehlens von Zusammenhängen zwischen den untersuchten Größen zeigten folgendes:
    a) in den kontextuellen modalen Verhältnissen der zahlenstärksten Gruppe substantivischer Realisationen mit dem Modalverb *können* zum komplexen Ausdruck objektiver Modalität mit dem Inhalt einer Möglichkeit und Fähigkeit oder zum komplexen Ausdruck subjektiver Modalität mit dem Inhalt einer Ungewissheit zeigen die Realisationen monosemer Substantive ihre statistische Relevanz;
    b) in den kontextuellen modalen Verhältnissen der hinsichtlich ihrer Zahlenstärke zweiten Gruppe substantivischer Realisationen mit dem Modalverb *müssen* zum komplexen Ausdruck objektiver Modalität mit dem Inhalt einer Notwendigkeit oder zum komplexen Ausdruck subjektiver Modalität mit dem Inhalt einer Gewissheit und Überzeugung wurde kein statistisch relevanter Zusammenhang zwischen den substantivischen Realisationen und dem gegebenen Modalverb festgestellt;

c) in den kontextuellen modalen Verhältnissen der hinsichtlich ihrer Zahlencharakteristiken dritten Gruppe substantivischer Realisationen mit dem Modalverb *sollen* zum komplexen Ausdruck objektiver Modalität mit dem Inhalt eines Auftrags, einer Verpflichtung und Empfehlung oder zum komplexen Ausdruck subjektiver Modalität bei der Distanzierung des Sprechers von einer fremden Behauptung kontextuell nicht genannter Personengruppe über jemand oder etwas zeigen die Realisationen monosemer Substantive ihre statistische Relevanz;
d) in den kontextuellen modalen Verhältnissen der – hinsichtlich ihrer Zahlencharakteristiken – vierten Gruppe substantivischer Realisationen mit dem Modalverb *lassen* zum komplexen Modalitätsausdruck mit dem Inhalt einer Veranlassung (Anordnung, Beauftragung), einer Erlaubnis, einer Zulassung sowie einer Möglichkeit/ Unmöglichkeit, einer Tauglichkeit/Untauglichkeit sind die Realisationen des monosemen Nomen-Potentials sowie der direkten Bedeutungen des polysemen Nomen-Potentials statistisch relevant, und der Chi-Quadrat-Wert für monoseme Substantive in der Gruppe mit dem Modalverb *lassen* ist der beste Wert von allen in der vorliegenden Arbeit untersuchten Größen;
e) in den kontextuellen modalen Verhältnissen der – hinsichtlich ihrer Zahlencharakteristiken – fünften Gruppe substantivischer Realisationen mit dem Modalverb *wollen* zum komplexen Ausdruck objektiver Modalität mit dem Inhalt einer Absicht, eines festen Willens oder zum komplexen Ausdruck einer subjektiven Modalität bei der Distanzierung eines Sprechers von der Behauptung einer fremden Person bezüglich sich selbst zeigen die Realisationen sowohl monosemer Substantive als auch direkter Nominationen der polysemen Substantive ihre statistische bzw. kontextuelle Relevanz;
f) in den kontextuellen modalen Verhältnissen der – hinsichtlich ihrer Zahlencharakteristiken – sechsten Gruppe substantivischer Realisationen mit dem Modalverb *mögen* zum komplexen Ausdruck objektiver Modalität mit dem Inhalt eines höflichen Wunsches, einer Abneigung oder zum komplexen Ausdruck subjektiver Modalität mit dem Inhalt einer einräumenden Vermutung statistisch relevant sind die Realisationen der Substantive mit ihren direkten und abgeleiteten Nominationen. Von allen Realisationen abgeleiteter substantivischer Nominationen mit Modalverben zeigen die abgeleiteten Nominationen einen statistisch relevanten Zusammenhang nur mit dem Modalverb *mögen*;
g) in den kontextuellen Verhältnissen der siebten, zahlenmäßig kleinsten Gruppe substantivischer Realisationen mit dem Modalverb

dürfen zum komplexen Ausdruck objektiver Modalität mit dem Inhalt einer Erlaubnis, eines Verbots oder zum komplexen Ausdruck subjektiver Modalität mit dem Inhalt einer Wahrscheinlichkeit statistisch relevant sind die Realisationen monosemer Substantive.

Zusammenfassend lässt sich aus unserer Tabelle feststellen, dass in den Spalten mit Zahlen und statistischen Berechnungen zu realisierten Nominationstypen der Substantive mit den Modalverben von links und von rechts generell ein statistisch relevanter Zusammenhang zwischen den untersuchten Größen beobachtet wird, was folglich im Endergebnis für eine kontextuelle Aktualität der Realisationen direkter substantivischer Nominationen und monosemer Substantive zum Ausdruck eines bestimmten modalen Inhalts in den modernen deutschsprachigen Prosawerken sprechen müsste.

Somit hat unsere Hypothese ihre Bestätigung gefunden, dass bei der Stellungnahme des Sprechers zur Geltung des modalen Sachverhaltes in syntaktischen Kontexten mit Modalverben die Realisationen monosemer Substantive sowie Realisationen der Substantive mit ihren direkten Nominationen tendenziell selektiert werden, um den komplexen modalen Inhalt explizit auszudrücken und somit die Intentionen des Textautors „ohne Lupe“ zu erschließen.

Selbstverständlich haben unsere Zusammenfassungen keinen Anspruch auf ein endgültiges Ergebnis, vielmehr geben sie einen Anstoß zu weiteren und tieferen Untersuchungen dieser Zusammenhänge nicht nur in den deutschsprachigen Prosawerken, sondern auch in den publizistischen Texten der deutschsprachigen Presse, weil es möglich macht, mithilfe von statistischen Berechnungen das ganzheitliche Bild in der Dynamik der Realisationszusammenhänge der Substantive mit den Modalverben zum Ausdruck des modalen Inhalts in der modernen deutschen Sprache möglichst objektiv zu studieren.

LITERATUR

Birkengof, G.M., Romm, Z.M., Uroeva, R.M. (1980). *Kurs grammatiki nemeckogo jazyka s grammatiko-fonetičeskimi upražnenijami. Časť 2*. Moskva: Meždunarodnye otnošenija.

Drebet V. (2006). Realizacija prjamyh ta pohidnyh nominacij imennykiv u prostyh ta skladnyh rečennjah nimeckoji presy. *Naukovyj visnyk Černiveckoho universytetu. Vyp. 319-320: Hermanska filolohija, 3-11.*

Drebet V. (2007). Triada “prjama/pohidna nominacija – proste/skladne rečennja – proste/skladne slovo” u nimeckij presi. *Naukovyj visnyk Černiveckoho universytetu. Vyp. 339-340: Hermanska filolohija, 20-38.*

Drebet V. (2008). Semantyčni hlybynnosti imennyka ta skladnopidrjadni hlybynnosti syntaksyčnyh konstrukcij u nimeckomovnyh prozovyh tvorah.

Problemy zagal´nogo, hermans´kogo ta slov´jans´kogo movoznavstva do 70-riččja professora V.V. Levic'koho. 194-202.

Drebet V. (2010). Imennyky u syntaksyčnyh ramkah pasyvnyh konstrukcij sučasnoji nimec'komovnoji presy. *Naukovyj visnyk Černivec'koho universytetu. Vyp. 531: Hermans'ka filolohija. 23-41.*

DUDEN (1999). *Das große Wörterbuch der deutschen Sprache: in zehn Bänden*/ hrsg. vom Wissenschaftlichen Rat der Dudenredaktion [Red. Bear.: Werner Scholze-Stubenrecht (Projektleiter) ... Unter Mitarb. von Brigitte Alsleben ...]. - 3., völlig neu bearb. Auflage. – Mannheim; Leipzig; Wien; Zürich: Dudenverlag.

Gak, V.G. (1969). *K probleme sintaksičeskoj semantiki (Semantičeskaja interpretacija „glubinnyh" i „poverhnostnyh" struktur. Invariantnyje sintaksičeskije značenija i struktura predloženija*. Moskva: Nauka.

Gak, V.G. (1973). Vyskazyvanije i situacija. In: *Problemy strukturnoj lingvistiki*. Moskva: Nauka.

Helbig G., Buscha J. (1996). *Deutsche Grammatik*. Leipzig; Berlin; München; Wien; Zürich; New York: Langenscheidt Verlag Enzyklopädie.

Jermolenko S.Ja. (1982). *Syntaksys i stylistyčna semantyka*. Kyiv: Naukova dumka.

Köhler R., Altmann G., Piotrowski R.G. (eds.) (2005). *Quantitative Linguistik. Ein internationales Handbuch*. Berlin/New York: de Gruyter.

Kozlovs´kyj V. (2008). Semantičnyj i prahmatičnyj aspekty rečennja (na materiali sučasnoji nimec'koji movy). In: *Problemy zagal'noho, hermans'koho ta slov'jans'koho movoznavstva. Do 70-riččja profesora V.V. Levic'koho. 323-326.*

Levyckij V.V. (2006). *Semasiolohija*. Vinnycja: Nova Knyga.

Metzler-Lexikon Sprache (1993). *Metzler-Lexikon Sprache*/hrsg. von Helmut Glück. Stuttgart; Weimar: Verlag J.B. Metzler.

Polenz P. (1988). *Deutsche Satzsemantik*. Berlin/New York: Walter de Gryeter.

Rüschel U. (1975). *Semantisch-syntaktische Relationen.* Max Niemeyer Verlag: Tübingen.

Šendel´s E.I. (1988). *Praktičeskaja grammatika nemeckogo jazyka*. Moskva: Vysšaja škola.

Tjagil', I.P. (2002). *Grammatika nemeckogo jazyka*. Sankt-Peterburg: Karo.

Zöfel P. (1992). *Statistik in der Praxis*. UTB, Stuttgart.

Texte

ChW = Wolf C. (1988). *Kindheitsmuster*. Frankfurt am Main: Luchterhand Literaturverlag.

HB = Böll H. (2007). *Billard um halb zehn.* Sankt-Peterburg: Karo.

HH = Hesse H. (2006). *Das Glasperlenspiel.* Sankt-Peterburg: Karo.

MW = Walser M. (2006). *Der Augenblick der Liebe*. Hamburg: Rowohlt Taschenbuch Verlag.
PS = Süskind P. (1994). *Das Parfum*. Zürich: Diogenes.

Kelih, E. et al. (eds.), Issues in Quantitative Linguistics, Vol. 2, 57-74

Vergleichende Analyse der polysemen Mikrostrukturen von *Land – Volk – Staat* und *zemlia – narod – derzhava* in der deutschen und der ukrainischen Presse

Sergej Kantemir

1. Ziele

Die Typologie der Polysemie in verschiedenen Sprachen wurde in vielen Abhandlungen von ukrainischen und europäischen Linguisten untersucht (Kijko, Levickij 2005; Kočerhan 2004; Levickij 2003; Legurska, Beceva 2003; Manakin 2004; Sandhop 2003; Schafikov 2004 u.v.a.). Im Vergleich zu den offensichtlichen Errungenschaften der kontrastiven Linguistik in diesem Bereich gilt es allerdings festzustellen, dass die Untersuchungen in der vergleichenden lexikalischen Semantik immer noch über keine exakten Verfahrensweisen sowie festgelegten wissenschaftlichen Termini verfügen (Kočerhan 1996: 4). In der Tat soll jeder Analyse interlingualer Korrelationen der systematische Ansatz zugrunde liegen, bei dem alle Beziehungen und Verhältnisse zwischen lexikalischen Einheiten jeder Vergleichssprache – paradigmatische, syntagmatische etc. – in Betracht gezogen werden.

Als Hauptkategorie der kontrastiven Studien gilt vor allem der Wert der sprachlichen Einheit, der durch deren Stelle in einem Sprachsystem bestimmt wird. Darum sind wir der Ansicht, dass es viel effizienter wäre auf der interlingualen Ebene nicht einzelne, isolierte Lexeme, sondern kleinere lexikalisch-semantische Paradigmen zu untersuchen. So lassen sich beispielsweise durch quantitative Differenzen und Ähnlichkeiten von Bestandteilen der Vergleichsmikrostrukturen anhand verschiedener statistischer Ansätze qualitative Besonderheiten dieser lexikalischen Teilgebiete feststellen. Durch die Verbindungen zwischen Wortfeldern, wie ja schon bekannt ist, offenbart sich dem Forscher das ganze lexikalisch-semantische System, also auch das sprachliche Weltbild (Kočerhan 1996: 5).

Allseitige grundlegende Änderungen im Leben einer Gesellschaft führen unvermeidlich zu regelmäßigen Transformationen im lexikalisch-semantischen Sprachsystem. Deshalb befinden sich die Studien zu Erneuerungstendenzen der gesellschaftlich-politischen Subsprache seit vielen Jahren im Fokus der sozialen Aufmerksamkeit (Karpilovska 2009: 127).

Die wichtigste Position in der sprachlichen Repräsentation von ideologischen Begriffen und Konzepten nehmen die politologische Lexik und Terminologie ein. Hier ist auch der Einstellungswandel im Bereich der gesellschaft-

lich-politischen Realitäten zu beobachten, was ihren Inhaltswert ändert, folglich lassen sich die Nominationen mancher Bezeichnungen variieren.

Das sprachliche Bild eines neuen Staates (wir sprechen in diesem Fall von der Ukraine) wird heute sowohl durch einheimische, als auch durch ausländische Massenmedien und Verlage gestaltet. Für die junge unabhängige Ukraine gelten *der Staat, die Gewalt* und *deren Institutionen* als grundlegende Begriffe für die Herausbildung eines neuen gesellschaftlichen Denkens.

In dieser Hinsicht halten wir für zweckmäßig, anhand der deutschen und ukrainischen Zeitungen eine kontrastive Untersuchung der Semantik und Funktion von zwei Mikrostrukturen *Land – Volk – Staat* und *земля – народ – держава* durchzuführen, die durch mehrdeutige Nomina mit spezifischem semantischem Inhalt vertreten sind. Bei dieser Untersuchung handelt es sich um die Analyse der Polysemie auf der interlingualen Ebene in den Feldstrukturen verschiedener Art.

Aufgabe des vorliegenden Beitrags wird es sein, mithilfe statistischer Ansätze die semasiologischen Eigenschaften der mehrdeutigen Substantive innerhalb von zwei Mikrostrukturen zu erschließen, die die zwischensprachlichen Korrelate *Land – земля, Volk – народ* und *Staat – держава* enthalten. Bestimmt können sich diese Polysemanten in den beiden Sprachen durch eigene verzweigte lexikalisch-semantische Struktur und ihre gemeinsamen inhaltlichen Schnittpunkte sowie unter dem Einfluss von außersprachlichen Faktoren zu gewissen Paradigmen vereinigen. Allerdings sind das Wesen dieser Beziehungen und der Grad der semantischen Affinität aufgrund ihrer Kombinierbarkeit mit syntagmatischen Partnern im Text noch gar nicht festgestellt geblieben. Wichtig ist auch zu erschließen, in welchen Bedeutungen die zu erforschenden lexikalischen Einheiten im modernen Zeitungsdiskurs zur Geltung kommen. Ferner steht es zu hoffen, dass die Präzisierung der Semantik dieser politologischen Termini im kontrastiven Aspekt zur Erneuerung des gesellschaftlich-politischen Wortschatzes der unabhängigen Ukraine beiträgt.

2. Materialien

Als Forschungsgegenstand diente eine Stichprobe, die im September-Oktober 2009 anhand der Tageszeitungen „Süddeutsche Zeitung“ (SZ) und „Україна молода“ (UM) (insgesamt eine halbe Million lexikalischer Einheiten) durchgeführt worden ist. Der relative Fehler der Stichprobe δ, der 5% beträgt, liegt im Bereich der statistischen Gesetzmäßigkeiten (Levickij 2004). Darüber hinaus wurden die Wörterbücher der deutschen und der ukrainischen Sprache (Duden - Deutsches Universalwörterbuch, 11-bändiges Wörterbuch der ukrainischen Sprache) zur Erschließung der lexikalisch-semantischen Struktur der erwähnten Polysemanten, deren Ähnlichkeiten und Unterschiede verwendet.

3. Bestandsaufnahme der mehrdeutigen Substantive

Bevor die Gebrauchshäufigkeit von polysemen Nomina in Texten untersucht werden konnte, wurden 600 Zeitungsartikel (je hundert für jedes Substantiv) in den o. a. Massenmedien bearbeitet und ein Register von Wörtern erstellt, das einer weiteren statistischen Analyse unterzogen wird (s. Tabelle 1).

Anhand dieser Tabelle wird deutlich, dass die Frequenzen der interlingualen Korrelaten prozentual fast völlig übereinstimmen, z.B.: *Staat* (19,2%) – *держава* (19,6%); *Volk* (12,5%) – *народ* (15,4%). Aber bei *Land – земля* wird festgestellt, dass die funktionelle Aktivität des Wortes *Land* in deutschen Zeitungstexten fast anderthalb mal größer ist, als der Gebrauch des Wortes *земля* in der Zeitung „Україна молода“ (vgl. 23,3% gegen 10%). Wahrscheinlich haben diese erheblichen Disproportionen sowohl einen außersprachlichen, als auch einen innensprachlichen Kontext.

Tabelle 1
Die quantitativen Charakteristika der untersuchten Nomina

Deutsche Nomina	Gebrauchshäufigkeit in Texten, %		Ukrainische Nomina	Gebrauchshäufigkeit in Texten, %	
Land	366	23,3%	*земля*	157	10%
Volk	196	12,5%	*народ*	244	15,4%
Staat	302	19,2%	*держава*	308	19,6%

Zum einen ist das mehrdeutige Substantiv *Land* durch eine außerordentliche Aktivität im gesellschaftlich-politischen Bereich geprägt, insbesondere zur Bezeichnung solcher Realien wie [Staat] und [Bundesland], in 366 Belegen wird das Wort *Land* 330-mal gerade in diesen Bedeutungen gebraucht. Und zum anderen wird der vom Lexem *Land* abgedeckter Denotatenbereich „Staat – Bundesland“ jedoch im Ukrainischen unter zwei Substantive *земля* und *країна* aufgeteilt, wobei die Bedeutung [Staat] nur in ca. 26% aller Kontexte mit dem Wort *земля* aktualisiert war.

4. Vergleichende Komponentenanalyse der lexikalisch-semantischen Struktur der polysemen Substantive

Die lexikalischen Einheiten zur Bezeichnung solcher Begriffe wie Staat, Volk, Land gehören normalerweise in jeder Sprache zu den gebräuchlichsten im gesellschaftlich-politischen Leben und werden dadurch oft zum beliebten Manipulationsobjekt seitens der Vertreter verschiedener Sprachgemeinschaften (Politiker, Reporter, Sprachwissenschaftler) (Agamben 2004).

Eine derartige Interpretation der oben erwähnten Wörter lässt uns eine sorgfältigere Analyse des semantischen Inhalts dieser Lexeme durchführen. Unter Hinzuziehung von lexikographischen Quellen können wir feststellen, welches Bedeutungsspektrum den Lexemen *Land* / *земля, Volk* / *народ, Staat* / *держава* eigen ist. Nach dem DUDEN (1996) bedeutet das Wort *Land*: 1. [Festland]; 2. [Ackerboden]; 3. [Gegend/dörfliche Gegend]; 4. [Staat]; 5. [Bundesland]; 6. [alle Bewohner eines Landes]. Man sieht, dass einige Bedeutungen ($Land_4$, $Land_6$) durch inhaltliche Korrelationen an die Lexeme *Staat* und *Volk* gebunden sein können.

Bei dem Wort *Volk* wurden fünf Bedeutungen festgestellt: 1. [Menschen mit gemeinsamer Geschichte, Sprache und Kultur]; 2. [Bevölkerung eines Landes, eines Staates]; 3. [die untere Schicht der Bevölkerung]; 4. [Menschenmenge]; 5. [Gemeinschaft von Insekten].

Die lexikalisch-semantische Struktur von *Staat* besteht aus folgenden Sememen: 1. [ein Land als politisches System]; 2. [Regierung]; 3. [Staatsgebiet, Kanton]; 4. [Bund, Bundesstaat]; 5. [Gemeinschaft von Tieren]; 6. [Pracht]; 7. [festliche Kleidung]. Es sei hervorgehoben, dass die Sememe $Staat_6$ und $Staat_7$ fast außer Gebrauch gekommen sind, was auch Spezialmarker im Wörterbuch bezeugen; darum werden in unserer Studie nur die ersten fünf Bedeutungen berücksichtigt. Bemerkenswert ist, dass sich die Semantik der Substantive *Staat* und *Volk*, durch die im Allgemeinen [Gemeinschaften von Menschen] bezeichnet werden, nach ihren denotativen Charakteristika etwas breiter erweist als die bei den ukrainischen Äquivalenten; so können einige Sememe (z.B. $Volk_5$, $Staat_5$) Gemeinschaften von Tieren bzw. Insekten bedeuten, dagegen sind diese Varietäten in den semantischen Strukturen von ukr. *народ* und *держава* nicht zu beobachten.

So können wir aufgrund der Komponentenanalyse der lexikalisch-semantischen Strukturen von mehrdeutigen Nomina *Land, Volk* und *Staat* einige inhaltliche Parallelitäten zwischen einzelnen Sememen innerhalb der untersuchten Mikrostruktur festhalten: $Land_4 – Staat_1$; $Land_5 – Staat_3$; $Land_6 – Volk_2$; sowie $Volk_5 – Staat_5$. Allerdings lassen sich der Charakter und die Größe derartiger Beziehungen nur anhand der Komponentenanalyse der in den lexikographischen Quellen festgelegten Wörter kaum erschließen; dafür werden auch Auszüge aus der „Süddeutschen Zeitung“ verwertet werden. Es ist jedoch höchstunwahrscheinlich, dass die betroffenen Wörter *Volk* und *Staat* in den sozial-politischen Texten die Bedeutung [Gemeinschaft von Insekten bzw. Tieren] zum Ausdruck bringen.

Und nun ermitteln wir die Semantik jeder lexikalischen Einheit innerhalb der Mikrostruktur *„земля – народ – держава“.* Im elfbändigen Wörterbuch der ukrainischen Sprache (SUM 1972, Bd. III: 557-559) findet man folgende Bedeutungen des Lexems *земля*: 1. [dritter Planet unseres Sonnensystems] ‘Planet Erde’; 2. [obere Schicht der Erdkruste] ‘Erdoberfläche’; 3. [dunkelbrauner Stoff als Teil der Erdoberfläche] ‘Lehm’; 4. [der aus festem Boden bestehende Teil der

Erdoberfläche (im Gegensatz zum Meer)] ‘Festland’; 5. [für den Anbau von Nutzpflanzen bestimmte Bodenfläche] ‘Ackerboden’; 6. [Land, Heimat, Staat].

Aufgrund des Vergleichs der lexikalisch-semantischen Strukturen von *Land* und *земля* lassen sich deren inhaltlichen Überschneidungen, wie in Abbildung 1 gezeigt, darstellen.

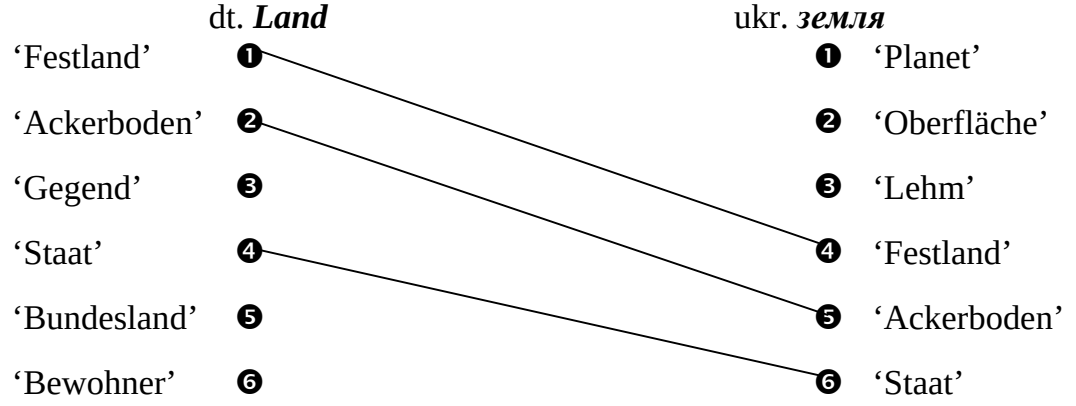


Abb. 1. Die strukturelle Stratifikation der Sememe in den Wörtern *Land* und *земля.*

Wie dieses Schema besagt, zeigen die analysierten Polysemanten anhand von einigen Sememen eine ausgeprägte Äquivalenz, wie z.B: 1) $Land_1$ – $земля_4$ mit der Bedeutung ‘Festland’; 2) $Land_2$ – $земля_5$ zur Bezeichnung eines Ackerbodens, sowie 3) $Land_4$ – $земля_6$ mit der Bedeutung ‘Staat’. Allerdings finden wir zu ukr. *земля* keine semantischen Korrelationen des Semems $Land_5$ mit der national-spezifischen Bedeutung einer territorialen Verwaltungseinheit [Gliedstaat eines Bundeslandes]. Bemerkenswert ist, dass die ersten drei Varietäten von *земля* auch keine Entsprechungen mit *Land* haben, weil sie semantisch an das andere polyseme deutsche Substantiv *Erde* gebunden sind, das alle erwähnten Varietäten vollständig involviert (vgl. *Die Erde ist unser Planet* ‘Земля – наша планета’; *die Erdfläche* ‘(земна) поверхня’; *die Heilerde* ‘лікувальна глина’). Was andere inhaltliche Analogien betrifft, so befindet sich einerseits das Verb mit ein und derselben Wurzel *landen* im Deutschen nach seiner Semstruktur am nächsten zu $земля_2$, andererseits konnte die übertragene Bedeutung $Land_6$ ‘Bewohner’ bei ukr. *земля* wahrscheinlich verloren gegangen sein bzw. durch das andere mehrdeutige Substantiv *країна* verdrängt worden sein, das nach seinen inhaltlichen Charakteristika die semantische Kette *земля – народ – держава* auch zu ergänzen ist (s. auch SUM 1973, Bd. IV: 320).

Das Lexem ukr. *народ* hat vier Bedeutungen, die mit den semantischen Komponenten von dt. *Volk* fast völlig identisch sind: 1. [Bevölkerung eines Landes, Bewohner eines Staates] ‘Bevölkerung’; 2. [Form der nationalen und ethnischen Einheit] ‘Nation’; 3. [Werktätige in der kapitalistischen Gesellschaft,

die untere Schicht der Bevölkerung] ‘Werktätige’; 4. [Menschen in großen Mengen] ‘Menschenmenge’ (SUM 1974, Bd. V: 174) (vgl. Abb. 2).

Die Tatsache, dass bei ukr. *народ* als Hauptbedeutung die Varietät ‘Bevölkerung eines Landes, Bewohner eines Staates’ anstatt ‘Nation’ im Wörterbuch auf Platz 1 angeführt worden ist, zeugt eher von der stark ausgeprägten Ideologisierung der damaligen Gesellschaft während der Erstellung des erklärenden Wörterbuches der ukrainischen Sprache; leider sind viele Ukrainer die überholte stalinistische Ideologie der ehemaligen Sowjetunion immer noch nicht losgeworden (s. auch Scheremeta 2009).

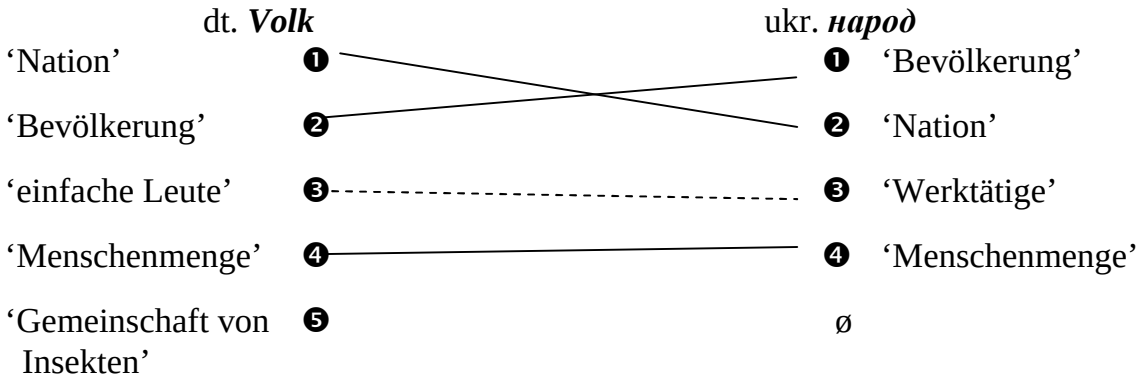


Abb. 2. Die semantischen Verbindungen zwischen *Volk* und *народ.*

Für das Wort *держава* findet man zwei Definitionen: *держава 1* als „1. ‘Gutsbesitz, Landgut’ (*veraltet*); 2. ‘Stärke, Festigkeit’ 3. ‘Führung’“ und *держава 2* mit den Bedeutungen: 1. [Apparat politischer Macht in der Gesellschaft] ‘Macht, Gewalt’ sowie 2. [Land mit dem Apparat politischer Macht] ‘Land, Staat’ (SUM 1971, Bd.II: 248). Indem wir die Analyse der gesellschaftlich-politischen Lexik durchführen, wird nur die zweite Definition in unserer Untersuchung berücksichtigt (vgl. Abb. 3).

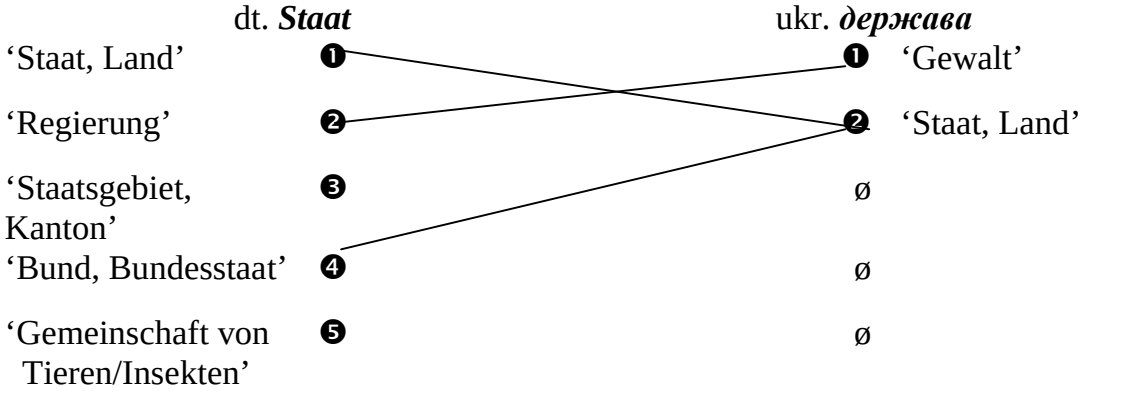


Abb. 3. Die strukturelle Stratifikation der Sememe in den Wörtern *Staat* und *держава.*

Bei dem Vergleich von dt. *Staat* und ukr. *держава* lässt sich eine unterschiedliche Bedeutungshierarchie feststellen (s. Abb. 3): das Semem *держава*$_1$ entspricht der zweiten übertragenen Bedeutung *Staat*$_2$, dabei korreliert *держава*$_2$ mit zwei Sememen *Staat*$_1$ und *Staat*$_4$. Dennoch gilt für die deutsche Sprache die semantische Entwicklung von *Staat*$_3$ und *Staat*$_4$ als spezifische Erscheinung, weil diese Sememe Polaritätsbedeutungen der Gattungs-Art-Verhältnisse erhalten haben, was im Ukrainischen nicht vorhanden ist. Anders gesagt, kann mithilfe des polysemen Substantivs „Staat" sowohl der Begriff „Teil von etwas" (*Gliedstaat*), als auch „etwas Ganzes" (*Bundesstaat*) verbalisiert werden, z.B.: *der US-**Staat** Texas – beide **Staaten** Deutschland und die USA.*

Im Allgemeinen lässt sich innerhalb der semantischen Kette *„земля – народ – держава"* nur eine inhaltliche Überschneidung *земля*$_6$ *– держава*$_2$ zur Bezeichnung eines Landes bzw. eines Staates feststellen.

5. Die funktionelle Besonderheiten der Polysemanten im Zeitungsdiskurs

Nach der Analyse der Gebrauchshäufigkeit der lexikalisch-semantischen Varianten von *Land, Volk, Staat, земля, народ, держава* wurde ermittelt, welche Bedeutungen in der funktionellen Hinsicht im Zeitungsdiskurs der Gegenwart am aktivsten sind (vgl. Tabelle 2).

Tabelle 2
Die quantitativen Charakteristika der Sememe von *Land*

Semem	Bedeutung	Anzahl
Land$_1$	[Festland]	10
Land$_2$	[Ackerboden]	5
Land$_3$	[Gegend]	12
Land$_4$	[Staat]	204
Land$_5$	[Bundesland]	126
Land$_6$	[alle Bewohner]	9
Insgesamt		366

Bei *Land* verteilt sich beispielsweise der größte Teil der Belege (90,2%) aus der „Süddeutschen Zeitung" auf die Bedeutungen [Staat] und [Bundesland]. Es ist ja auch kein Zufall, weil die gesellschaftlich-politische Lexik, die durch den publizistischen funktionellen Stil bedingt ist, in der Presse als dominierend gilt. Die anderen übertragenen Bedeutungen von *Land* haben dabei nur noch unwesentliche Gebrauchshäufigkeit, denn die Bedeutungsdiversifikation verläuft nach Diversifikationsgesetzen, daher müssen einige Sememe seltener vorkommen.

Beispiele:

Semem ***Land*$_1$**: *Kombiniert wird diese Politik mit einem globalen Infrastrukturprogramm – Häfen, Transportwege auf dem* ***Land****, Pipelines.* (SZ, 14.11.2009).

Semem ***Land*$_2$**: *Das* ***Land*** *wurde vielfach für lächerlich niedrige Summen verkauft, zum allergrößten Teil an Westdeutsche.* (SZ, 11.11.2009).

Semem ***Land*$_3$**: *Zaki ist ein alter Weggefährte des Präsidenten, und er hat ihn in diesen Tagen bei einer Reise durchs zerstückelte* ***Land*** *nach Hebron und Bethlehem begleitet.* (SZ, 11.11.2009).

Semem ***Land*$_4$**: *Ist Ihr* ***Land*** *inzwischen eine Demokratur geworden?* (SZ, 06.11.2009).

Semem ***Land*$_5$**: *Gemeinsam mit dem* ***Land*** *Bayern wird gerade ein Strukturpaket geschnürt.* (SZ, 15.11.2009).

Semem ***Land*$_6$**: *Rund 43 Prozent des* ***Landes*** *können von Hilfsorganisationen wegen der Gewalt nicht erreicht werden.* (SZ, 20.11.2009).

Tabelle 3
Die quantitativen Charakteristika der Sememe von *Volk*

Semem	Bedeutung	Anzahl
Volk$_1$	[Nation]	42
Volk$_2$	[Bevölkerung]	141
Volk$_3$	[untere Schicht]	6
Volk$_4$	[Leute]	7
Volk$_5$	[Gemeinschaft von Tieren]	-
Insgesamt		196

Tabelle 3 verdeutlicht u.a., dass die funktionelle Aktivität einzelner Bedeutungen des polysemen Substantivs *Volk* (196 Belege) auch einen differenziellen Charakter hat: am häufigsten wird *Volk* zur Bezeichnung der Bevölkerung eines Landes (141 Belege) und als Form der nationalen und ethnischen Einheit (42 Belege) gebraucht.

Beispiele:

Semem ***Volk*$_1$**: *Dem türkischen* ***Volk*** *selbst scheint das Thema weitestgehend egal zu sein.* (SZ, 13.10.2009).

Semem ***Volk*$_2$:** *In zwei Jahren wird es eine* ***Volks****zählung geben.* (SZ, 20.11.2009).

Semem ***Volk*$_3$:** *„Sie sehen nicht, was im* ***Volke*** *los ist“, sagte er in Richtung der Kanzlerin.* (SZ, 10.11.2009).

Semem ***Volk*$_4$:** *Das* ***Volk*** *jubelte ihm zu.* (SZ, 01.12.2009).

Wie schon vorausgesehen, wurden in der Zeitung keine Belege von *Volk* mit der Bedeutung [Gemeinschaft von Tieren / Insekten] nachgewiesen.

Tabelle 4
Die quantitativen Charakteristika der Sememe von *Staat*

Semem	Bedeutung	Anzahl
Staat$_1$	[Land]	196
Staat$_2$	[Regierung]	88
Staat$_3$	[Staatsgebiet]	15
Staat$_4$	[Bundesstaat]	3
Staat$_5$	[Gemeinschaft von Tieren]	-
Insgesamt		302

Genauso wurde das Substantiv *Staat* (302 Belege) nicht ein einziges Mal in der übertragenen Bedeutung [Gemeinschaft von Tieren / Insekten] gebraucht. Wie die Forschungsergebnisse bezeugen, bezeichnet die analysierte lexikalische Einheit am häufigsten (94%) die Begriffe [Land] und [Regierung]. Bemerkenswert ist, dass die Bedeutung von *Staat* als territoriale Verwaltungseinheit in allen Fällen zur Bezeichnung von US-Staaten realisiert wurde.

Beispiele:

Semem ***Staat*$_1$:** *Liberias Präsidentin … ist die einzige Frau an der Spitze eines afrikanischen* ***Staates.*** (SZ, 26.10.2009).

Semem ***Staat*$_2$:** *Im Zuge des Wirtschaftsaufschwungs übt* ***der Staat*** *seine Kontrolle nun anders aus.* (SZ, 13.10.2009).

Semem ***Staat*$_3$:** *Der Demokrat Bill Owens gewann eine Nachwahl zum Kongress im US-****Staat*** *New York.* (SZ, 04.11.2009).

Semem ***Staat*$_4$:** *US-Abgeordnete wollen Guantanamo-Insassen in* ***die Staaten*** *verlegen.* (SZ, 16.10.2009).

Wie die Varietätsanalyse aufgezeigt hat, verteilt sich die Funktionalität der übertragenen Bedeutungen von ukr. *земля* im Zeitungsdiskurs im Allgemeinen auf vier Sememe: am häufigsten wird das Substantiv *земля* in der Bedeutung [Ackerboden] (63 Belege) gebraucht, dann kommen die Bezeichnungen eines Landes / eines Staates (41 Belege) und der geografischen Begriffe (*Planet*, *Oberfläche*).

Beispiele:

Semem ***земля*$_1$:** *Руденко доводить, що не лише* ***Земля****, а й 80-90% нашої Метагалактики – живі* (UM, 19.12.2009). ‘Herr Rudenko behauptet, dass nicht nur die Erde, sondern auch 80-90% unserer Metagalaxie lebendig sind’.

Semem ***земля*$_2$:** *Москву готові стерти з лиця* ***землі*** *через поклади діамантів під нею* (UM, 26.12.2009). ‘Man ist bereit, Moskau wegen Diamantenvorkommen unter der Stadt dem Erdboden gleichzumachen’.

Semem ***земля*$_3$:** *Поки ми їх шукаємо, думаю про те, що до цієї* ***землі*** *мусять прилипати не лише жіночі чобітки чи натруджені селянські руки, а й закохані у село душі* (UM, 23.12.2009). ‘Solange wir nach ihnen suchen,

denke ich daran, dass nicht nur die Damenstiefel oder die von der schwerer Arbeit erschöpften Bauernhände an diese Erde festkleben, sondern auch die ins Dorf verliebte Seelen'.

Semem ***земля$_4$*: *Сосни шумлять над ошатними доріжками, і в тебе виникає відчуття острів'янина, відірваного від метушні „великої **землі**“* (UM, 10.12.2009). 'Die Kiefer rascheln über den schönen Wegen, und es fühlt sich, als wärest du ein Insulaner, der von der Hektik des „Festlandes“ abgeschieden ist'.

Semem ***земля$_5$*: *Унікальні ландшафти поєднуються тут із такою ж унікальною рослинністю: на ніколи не ораних **землях** росте чимало реліктових рослин* (UM, 11.11.2009). 'Einmalige Landschaften verbinden sich mit der so einzigartigen Pflanzenwelt: Auf dem noch nie gepflügten Ackerland gedeihen viele Reliktpflanzen'.

Semem ***земля$_6$*: *Вона залишила нам імена героїв, які своїм життям не раз доводили, що український патріотизм живий і буде жити вічно, поки є українська **земля** та український народ* (UM, 29.12.2009). 'Sie hat uns die Namen von Helden bewahrt, die mit ihrem Leben mehrmals bewiesen haben, dass der ukrainische Patriotismus lebendig ist und so bleibt, solange es das Land Ukraine und das ukrainische Volk gibt'.

Tabelle 5
Die quantitativen Charakteristika der Sememe von *земля*

Semem	Bedeutung	Anzahl
Земля$_1$	[Planet]	27
земля$_2$	[Oberfläche]	24
земля$_3$	[Lehm]	1
земля$_4$	[Festland]	1
земля$_5$	[Ackerland]	63
земля$_6$	[Staat]	41
Insgesamt		157

Bemerkenswert ist, dass das Substantiv *народ* in den modernen ukrainischen Zeitungstexten 140-mal (71,4%) in der Bedeutung [Nation] gebraucht wird, was voraussichtlich vom besonderen staatsaufbauenden Wert dieses Begriffs zeugt. Als zweithäufigstes Sem wird *народ$_1$* mit der Bedeutung [Bevölkerung eines Landes] festgestellt. Wenn wir die Gebräuchlichkeit mancher Varietäten von dt. *Volk* in der „Süddeutschen Zeitung“ (s. Tabelle 3) vergleichen, so ist dabei die Aktualisierung des Sems [Nation] in den Hintergrund geraten, indem das Sem [Bevölkerung] immer mehr an Bedeutung gewinnt.

Tabelle 6
Die quantitativen Charakteristika der Sememe von ukr. *народ*

Semem	Bedeutung	Anzahl
народ$_1$	[Bevölkerung]	43
народ$_2$	[Nation]	140
народ$_3$	[Werktätige / einfache Leute]	6
народ$_4$	[Menschenmenge]	7
Insgesamt		196

Beispiele:

Semem ***народ*$_1$:** *…Сталін морив голодом власний* ***народ*** (UM, 14.11.2009). '…Stalin ließ das eigene Volk verhungern'.

Semem ***народ*$_2$:** *…Тільки коли здобудемо свою самостійну державу, український* ***народ*** *матиме можливість побудувати вільне національне і справедливе життя* (UM, 05.11.2009). '…Erst wenn wir einen eigenen selbstständigen Staat gewonnen haben, wird das ukrainische Volk die Möglichkeit bekommen, ein freies nationales und gerechtes Leben zu gestalten'.

Semem ***народ*$_3$:** *…Проігнорували не мене – проігнорували* ***народ****, людей* (UM, 05.11.2009). '…Nicht ich wurde ignoriert, sondern das Volk, einfache Leute wurden nicht beachtet'.

Semem ***народ*$_4$:** *…Спускалися з гори вниз автобусом, де весь* ***народ*** *чхав, кашляв* (UM, 07.11.2009). '…Talgefahren sind wir mit dem Bus, wo alle Menschen niesten und husteten'.

Es sei betont, dass eine relativ große Menge des gemeinsamen Vorkommens der beobachteten Substantive innerhalb ein und desselben Satzes auch von deren semantischen Überschneidungen und der Distributionsähnlichkeit zeugen kann, wie etwa: *І немає сили, що спинила б цілеспрямований поступ українського* ***народу*** *до своєї національної* ***держави*** (UM, 19.11.2009). 'Es gibt wohl keine Kraft mehr, die die zielbewusste Vorwärtsbewegung des ukrainischen Volkes zu seinem nationalen Staat aufhalten würde'.

Das Substantiv *держава* wird in meisten Fällen (93,8%) zur Bezeichnung eines Landes mit einem Apparat politischer Macht in der Gesellschaft gebraucht, dabei stellt man nur noch 19-mal die Bedeutung [Regierung, Gewalt, Institution] fest.

Tabelle 7
Die quantitativen Charakteristika der Sememe von *держава*

Semem	Bedeutung	Anzahl
держава$_1$	[Gewalt]	19
держава$_2$	[Land, Staat]	289
Insgesamt		308

Beispiele:

Semem ***держава₁:*** *…**держава** обраховує цих людей і закладає на них гроші в бюджет* (UM, 17.09.2009). '…die Regierung betrügt diese Leute und berechnet für sie das Geld im Haushalt [für das kommende Jahr]'.

Semem ***держава₂:*** *Минулого тижня у Нью-Йорку збиралися не лише лідери світових **держав** на 64-й сесії Генасамблеї ООН, а й зірки шоу-бізнесу, яким не байдужі глобальні проблеми* (UM, 29.09.2009). 'Vorige Woche sind nicht nur die Staatschefs aus der ganzen Welt zur 64. Tagung der UN-Generalversammlung in New York zusammengetreten, sondern auch die Popstars, denen die Globalprobleme nicht gleichgültig sind'.

Somit ergibt sich für die beiden Sprachen: Die höchstgebräuchlichste gemeinsame Bedeutung für die beiden Mikrostrukturen wurde auf der intersprachlichen Ebene durch den Begriff *eines Staates* realisiert.

6. Die Kombinierbarkeit der polysemen Substantive im Modell [Wort + Subklasse]

Zwischen der Wortbedeutung und derer Kombinierbarkeit besteht, wie bekannt, eine gewisse Korrelation. Einige Linguisten sind der Ansicht, die Bedeutung des Wortes sei seine potenzielle Kombinierbarkeit, d.h. je größer die eventuelle Kombinierbarkeit eines Wortes ist, desto breiter werden die neuen Tendenzen und Aspekte seines Gebrauchs in der Sprache (Zvegincev 1957: 132). Wenn wir allerdings annehmen, es bestehe zwischen den Wörtern, die in einem Text unter gleichen Distributionsbedingungen vorkommen, eine bestimmte semantische Beziehung, so lässt sich der Grad dieser Beziehung feststellen, nachdem wir quantitative Charakteristika der Kombinierbarkeit der zu erforschenden Substantive mit ihren kontextuellen Partnern durch spezielle statistische Verfahren erschlossen haben.

Zur Untersuchung der semantischen Kombinierbarkeit der intersprachlichen Korrelate *Land – земля, Volk – народ* und *Staat – держава* im Modell *[Wort + Wortsubklasse]* muss vorher ein Inventar von Subklassen aller Distributionselemente aufgestellt werden, die unter einem bestimmten semantischen Merkmal zusammengefasst sind.

Unter Hinzuziehung von vorangehenden Erfahrungen in der Linguistik (s. Levickij 1989: 135-136) wurde eine eigene semantische Klassifikation der syntagmatischen Partner erarbeitet, die entsprechend den Befunden aus der „Süddeutschen Zeitung" und der Tageszeitung „Україна молода" den Charakter kontextueller Umgebung der beobachteten polysemen Substantive in vollem Umfang darstellt.

In Anbetracht des kontrastiven Charakters der Studie wurden einige Verfahrensbedingungen bei der Aufstellung dieses Inventars zwecks der ausreichenden Objektivität angewandt:

1) in die Liste wurden sowohl Adjektive als auch Substantive aufgenommen;
2) das innere Syntagma in Zusammensetzungen wurde dem äußeren Syntagma als dasjenige gleichgesetzt, das in ein äußeres umgewandelt werden kann, z.B.: *die **Volk**spolizei > *die Polizei des **Volkes*** usw.
3) da das aufgestellte Register in gleichem Maße zu beiden Mikrostrukturen *Land – Volk – Staat* und *земля – народ – держава* gehört, lässt sich annehmen, dass dieses sog. *tertium comparation*is eine gewisse „metalinguistische" Funktion hat/habe.

Folglich sind alle kontextuellen Partner des aufgestellten Registers, die in Wortverbindungen mit den beobachteten Polysemanten fixiert sind, zu Subklassen (Clustern) wie folgt zusammengefasst worden**:**

1) Führungsstellen, Berufe: *der Chef, der Präsident; глава* 'der Chef, das Oberhaupt'*, урядовці* 'die Regierungsmitglieder';

2) Nationalität und geografische Begriffe: *deutsch, arabisch; український* 'ukrainisch'*;*

3) Ideologisch-rechtliche Begriffe, Attribute und Funktionen: *das Recht, die Macht, die Propaganda; політика* 'die Politik'*, конституція* 'die Verfassung'*;*

4) Gemeinschaften, Teil-Ganzes-Begriffe: *die EU-Zone, der Bund, der Teil; ЄС* 'EU'*;*

5) Finanzen, Wirtschaft: *die Industrie, die Krise; дотації* 'die Subventionen'*, баланс* 'die Bilanz';

6) Sicherheit: *die Sicherheit, die Grenze, der Krieg; безпека* 'die Sicherheit';

7) Institutionen, Organisationen: *das Parlament, die Regierung; база* 'die Station; die Zentralstelle; der Stützpunkt'*, парламент* 'das Parlament'*;*

8) Menschen: *der Nachbar, der Gast, die Mannschaft; партнер* 'der Partner'*, ворог* 'der Feind'*, населення* 'die Bevölkerung'*;*

9) Gegenstände: *der Wagen, die Zeitung, das Flugzeug; сателіт* 'der Satellit'*;*

10) Allgemeine Charakteristik: *frei, arm, exotisch; історичний* 'historisch, geschichtlich'*, великий* 'groß'*;*

11) Sonstige Begriffe: *der Abend, das Gefühl, der Trend; пуп* 'der Nabel; der Mittelpunkt' *,вчинки* 'die Handlungen, die Taten'*, психологія* 'die Psychologie'.

Die Gebrauchshäufigkeit aller 11 Subklassen der kontextuellen Partner mit den zu untersuchenden Substantiven ist in Tabelle 8 aufgeführt. Für den Gesamtüberblick über das Kombinierbarkeitspotenzial der analysierten Wörter wurden auch die Befunde in die Tabelle aufgenommen, wo diese allein (das Null-Syntagma) im Text vorkommen.

Tabelle 8
Die syntagmatischen Beziehungen der beobachteten Substantive

Stichprobe „SZ“				Kontextuelle Partner	Stichprobe „УМ“			
Σ	*Land*	*Volk*	*Staat*		*земля*	*народ*	*держава*	Σ
92	18	9	65	Führung, Berufe	0	1	42	43
86	33	18	35	Geografische Begriffe	61	61	68	190
134	7	79	48	Ideologische Begriffe	2	14	9	25
66	39	3	24	Gemeinschaften	13	1	8	22
59	21	10	28	Finanzen	12	4	5	19
35	6	17	12	Sicherheit	1	0	2	3
98	66	14	18	Institutionen	1	0	3	4
40	16	7	17	Menschen	7	6	6	19
22	4	15	3	Gegenstände	0	0	1	1
49	32	3	16	Allgemeine Charakteristik	26	32	74	132
18	15	2	1	Sonst. Begriffe	5	9	6	20
170	109	26	35	Null-Syntagma	29	117	84	230
864	366	196	302	Insgesamt	157	244	302	703

Wie aus Tabelle 8 ersichtlich, unterscheiden sich die Frequenzen der Kombinierbarkeit der Substantive in den untersuchten Mikrostrukturen erheblich voneinander:

1. Die Mikrostruktur *Land – Volk – Staat* ist durch einen größeren Kombinierbarkeitsumfang gekennzeichnet, der einen gleichmäßigeren unterteilten Charakter hat, als der in der Mikrostruktur *земля – народ – держава*.

2. Dass die Subklassen „Nationalität und geografische Begriffe“ und „Allgemeine Charakteristik“, die im Allgemeinen aus Adjektiven bestehen, am häufigsten mit den ukrainischen Substantiven gebraucht werden, zeugt wohl davon, dass es in den ukrainischen Zeitungstexten mehr attributive Verbindungen gibt als in den deutschen.

Um die Stufe der semantischen Verbindung zwischen Wörtern zu messen, wurde der χ^2-Test für einzelne Zellen angewandt (s. Levickij 1989; 2004).

Nach der durchgeführten statistischen Analyse wurde festgestellt, dass die syntagmatischen Verbindungen des Substantivs *Land* mit folgenden Subklassen der kontextuellen Partner statistisch signifikant verbunden sind: 1) *Land* + Subklasse „Institutionen“ (χ^2 = 28,26); 2) *Land* + Subklasse „Sonstige Begriffe“ (χ^2 = 12,64); 3) *Land* + Subklasse „Allgemeine Charakteristik“ (χ^2 = 9,22); 4) *Land* + Subklasse „Gemeinschaften, Teil-Ganzes-Begriffe“ (χ^2 = 8,19).

Für das Substantiv *Volk* wurden folgende statistisch signifikante Verbindungen festgestellt: 1) *Volk* + Subklasse „Ideologisch-rechtliche Begriffe“ (χ^2 =

118,96); 2) *Volk* + Subklasse „Gegenstände“ (χ^2 = 26,64); 3) *Volk* + Subklasse „Sicherheit“ (χ^2 = 13,94).

Das Wort *Staat* kommt in Verbindungen mit zwei Subklassen: 1) *Staat* + Subklasse „Führung, Berufe“ (χ^2 = 57,71); 2) *Staat* + Subklasse „Menschen“ (χ^2 = 4,14) signifikant oft vor.

Innerhalb der Mikrostruktur *земля – народ – держава* wurden sechs syntagmatischen Beziehungen mit ihren kontextuellen Partnern festgestellt, die eine statistische Signifikanz haben: 1) *земля* + Subklasse „Gemeinschaften, Teil-Ganzes-Begriffe“ (χ^2 = 17,98); 2) *земля* + Subklasse „Finanzen, Wirtschaft“ (χ^2 = 15,38); 3) *земля* + Subklasse „Nationalität und geografische Begriffe“ (χ^2 = 14,94); 4) *народ* + Subklasse „Ideologisch-rechtliche Begriffe“ (χ^2 = 5,35); 5) *держава* + Subklasse „Führung, Berufe“ (χ^2 = 54,80); 6) *держава* + Subklasse „Allgemeine Charakteristik“ (χ^2 = 10,51).

Bemerkenswert ist allerdings, dass auf der intersprachlichen Ebene eine vollständige Übereinstimmung der festgelegten Beziehungen ermittelt worden ist, nämlich: 1) die beiden Substantive dt. *Land* und ukr. *земля* haben statistisch signifikante Beziehungen mit ein und derselben Subklasse „Gemeinschaften, Teil-Ganzes-Begriffe“; 2) sowohl dt. *Volk*, als auch ukr. *народ* kommen in statistisch signifikanten Verbindungen mit der gemeinsamen Subklasse „Ideologisch-rechtliche Begriffe“ vor; 3) das intersprachliche Korrelat *Staat – держава* weist statistisch signifikante Beziehungen mit der Subklasse, die Führung und Berufe bezeichnen, auf. Darüber hinaus ist das Äquivalenzpaar *Land – держава* durch statistisch signifikante Verbindungen mit der Subklasse „Allgemeine Charakteristik“ gekennzeichnet. Diese Tatsache lässt die Annahme zu, dass man sie als stabile kontextuelle Partner bezeichnen kann. Die Beziehungen der übrigen Syntagmen haben keine statistische Signifikanz, was von ihren zufälligen Charakter zeugen mag.

7. Paradigmatische Relationen zwischen den Substantiven

Wie schon viele Sprachforscher gezeigt haben, können die paradigmatischen Beziehungen zwischen den Wörtern aufgrund der Eigenschaften ihrer Kombinierbarkeit behandelt werden (Moskovič 1969). Mit Hilfe der Korrelationsanalyse lassen sich die Kombinierbarkeitsmerkmale zweier Wörter paarweise berücksichtigen bzw. messen. Unter Hinzuziehung von der Hypothese, dass die Ähnlichkeit der lexikalischen Kombinierbarkeit von der semantischen Ähnlichkeit der Wörter zeugen müsste, haben wir eine solche Korrelationsanalyse auf Grund der Frequenzcharakteristika aus der Tab. 8 durchgeführt. Dabei wurden nur die Fälle berücksichtigt, wo die gesamte Gebrauchshäufigkeit der Nomina mit den kontextuellen Partnern mindestens 5 Belege betragen. Die Ergebnisse der Korrelationsanalyse sind in Tab. 9 angeführt.

Tabelle 9
Die paradigmatischen Relationen der beobachteten Substantive

	Land	*Volk*	*Staat*	*земля*	*народ*	*держава*
Land		- 0,06	0,14	0,38	0,75	0,55
Volk			0,43	- 0,08	0,16	- 0,04
Staat				0,14	0,22	0,37
земля					0,70	0,75
народ						0,83
держава						

Tabelle 9 verdeutlicht u.a., dass die Korrelationskoeffizienten sowohl positive (mit dem Zeichen "+"), als auch negative Werte haben. Das bedeutet, dass nicht alle beobachteten Substantive mit ihren kontextuellen Partnern gleichmäßig auftreten. Deshalb hat das Korrelationsverhältnis zwischen den ausgewählten lexikalischen Einheiten einen unsymmetrischen Mischcharakter (vgl. *Staat – Volk* (0,43), *Land – Volk* (-0,06) usw.).

Darüber hinaus wird anhand dieser Tabelle deutlich, dass die hohe Gebrauchshäufigkeit der untersuchten Lexeme nicht immer von deren festen paradigmatischen Beziehungen innerhalb ihrer Mikrosysteme zeugt.

Die Forschungsergebnisse besagen, dass die paradigmatischen Relationen zwischen den ukrainischen Substantiven in der Mikrostruktur *земля – народ – держава* durch einen viel höheren Korrelationsgrad ausgezeichnet werden, als diese zwischen ihren Äquivalenten in der deutschen Sprache. Gerade zwischen diesen Wortpaaren wurde die größte Verbindungsstärke – also die Ähnlichkeit der Distribution – festgestellt: *народ – держава* (0,83); *земля – держава* (0,75); *земля – народ* (0,70). Bei 12-2 = 10 Freiheitsgraden ist der kritische Werte des Korrelationskoeffizienten auf der 0.05-Ebene gleich 0.58 und auf der 0.01-Ebene gleich 0.71. Im Vergleich dazu werden die Substantive der semantischen Kette *Land – Volk – Staat* durch schwache und ungleichmäßige Relationen (*Land – Staat* (0,14); *Land – Volk* (-0,06)) verbunden, darunter belegt das Wortpaar *Volk – Staat* mit dem Spitzenwert 0,43 Platz 1.

Interessant ist allerdings, dass die Korrelationsanalyse ermöglicht, paradigmatische Beziehungen auch zwischen anderen intersprachlichen Korrelaten zu erschließen: So hat dt. *Land* sehr feste Relationen mit dem ukrainischen Nomen *народ* (0,75), sowie gemäßigte paradigmatische Beziehungen mit ukr. *держава* (0,55). Außerdem wurden fast gleiche Korrelationen bei den Wortpaaren *Land – земля* (0,38) i *Staat – держава* (0,37) festgestellt. Ein solches Korrelationsverhältnis kann sowohl von der gleichen Frequenz, als auch von der Ähnlichkeit der Distribution innerhalb des aufgestellten Registers der kontextuellen Partner (in unserem Fall – *tertium comparationis*) zeugen. Die übrigen intersprachlichen Korrelate haben keine statistische Signifikanz.

8. Schlussfolgerungen

Die durchgeführte Untersuchung lässt einige qualitative und quantitative Besonderheiten bei den semantischen Ketten *Land – Volk – Staat* und *земля – народ – держава* feststellen:

1. Am häufigsten wurden in den deutschen Zeitungstexten folgende Bedeutungen (Bedeutungsvarianten) gebraucht: 1) [Land, Staat] (400 Belege oder 56,4%), 2) [Bevölkerung eines Landes] (27,8%), 3) [Bundesgebiet] (17, 9%), 4) [Gewalt, Regierung] (12,5%), in der Mikrostruktur *земля – народ – держава* hatten die Sememe mit den Bedeutungen [Land, Staat] (330 Belege oder 38,2%), [Nation] (16,2%), 3) [Ackerland] (7,3%) und [Bevölkerung] (5%) die höchsten Frequenzwerte. Dabei wurde festgestellt, dass die gebräuchlichste gemeinsame Bedeutung für die beiden Mikrostrukturen auf der intersprachlichen Ebene durch den Begriff *eines Staates* realisiert wurde.

2. Die beiden Teilgebiete zeigen eine deutlich ausgeprägte Ähnlichkeit der Distribution aufgrund der Subklassen von Substantiven, die Führungsstellen und Berufe, ideologisch-rechtliche Begriffe, Gemeinschaften und Teil-Ganzes-Begriffe bezeichnen, sowie der Subklasse von Adjektiven der allgemeinen Charakteristik.

3. Die Substantive *Volk – народ* kommen in statistisch signifikante Verbindungen mit der gemeinsamen Subklasse „Ideologisch-rechtliche Begriffe“. Die Beziehungen von dt. *Land* und ukr. *земля* haben mit der ein und derselben Subklasse „Gemeinschaften und Teil-Ganzes-Begriffe“ auch statistische Signifikanz. Das intersprachliche Korrelat *Staat – держава* demonstriert dabei statistisch signifikante Relationen mit der Subklasse, die Führungsstellen und Berufe bezeichnen.

4. Die Anzahl der attributiven Kollokationen mit den untersuchten lexikalischen Einheiten ist in den ukrainischen Zeitungsartikeln viel größer, als die mit den Substantiven *Land, Volk, Staat* in der „Süddeutschen Zeitung“.

5. Die paradigmatischen Relationen zwischen den Substantiven in der Mikrostruktur *земля – народ – держава* zeichnen sich durch einen relativ höheren Korrelationsgrad als die Beziehungen zwischen ihren Äquivalenten in der deutschen Sprache aus, auf der intersprachlichen Ebene wurden dabei gemäßigte sowie im Großen und Ganzen schwache statistisch signifikante Korrelationen zwischen den beobachteten Wörtern festgestellt.

Es wäre auch interessant, den Grad der semantischen Ähnlichkeit anhand von anderen Mikrostrukturen in den ukrainischen und deutschen Zeitungstexten weiter zu untersuchen.

Literatur

Agamben G. (2004). What is a People? http://makeworlds.net/node/108

Duden. (1996). *Deutsches Universalwörterbuch* (Hrsg. von G. Drosdowski). Mannheim-Wien-Zürich: Duden-Verlag.

Karpilovska, E. (2009). Obraz derzavy u movi vidkrytoho suspilstva: novomova čy mova novoho myslennja? In: *Movy ta kultury u novij Evropi: 127-138.* Kyjiw: Kyjiwskyj universitet.

Kijko, S., Levickij, **V.** (2005). Statystyčni doslidzennja polisemiji dijesliv sučasnoji nimeckoji movy. In: *Problemy kvantytatyvnoji linguistyky: 210-244*. Černivci, Ruta.

Kočerhan, M. (1996). Zistavna leksyčna semantyka: problemy i metody doslidzennja. *Movoznavstvo 2-3, 3-12.*

Kočerhan, M. (2004). Zistavne movoznavstvo i problema movnyx kartyn svitu. *Movoznavstvo 2-3, 3-12.*

Legurska, P., Bečeva, N. (2003). Nekotoryje problemy sopostavitelno-tipologičeskogo analiza predmetnyx imen v russkom, serbskom i bolgarskom jazykax. *Srpski jezik 8(1-2), 279-290.*

Levickij, V. (2003). Leksyčna polisemija ta kvantytatyvni metody jiji doslidzennja. *Movoznavstvo 4, 17-25.*

Levickij, V. (2004). *Kvantytatyvni metody v linguistike.* Černivci: Ruta.

Levickij, V.V. (1989). *Statističeskoe izučenie lexičeskoj semantiki*. Kiev: Minvuz.

Manakin, V. (2004). *Sopostavitelnaja leksikologija*. Kiev: Znannja.

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

Sandhop, M. (2003). Von *Abend* bis *Zunge*: Lexikalische Semantik des Deutschen, Tschechischen, Englischen und Französischen im Vergleich. Frankfurt am Main: Peter Lang.

Schafikov, S. (2004). *Tipologija leksičeskix sistem i leksiko-semantičeskix universalij*. Ufa: Baschkirskij universitet.

Scheremeta, N. (2009). Psyxolohičnyj linguocyd v Ukrajini. Naukovi praci Kamjanec-Podilskoho nacionalnoho universytetu. *Filolohični nauky 20, 777-779.*

SUM (1970-1980). *Slovnyk ukrajinskoji movy*. In 11 Bänden. Kyjiw: Naukova dumka.

Zvegincev, V.A. (1957). *Semasiologija*. Moskva: Moskovskij universitet.

Kelih, E. et al. (eds.), Issues in Quantitative Linguistics, Vol. 2, 75-87

Frequency structure of New Year's presidential speeches in Czech. The authorship analysis

Radek Čech

1. Introduction

New Year's presidential speeches represent a very specific genre. They mix both political and festive aspects and, contrary to common political speeches, they usually do not have persuasive character. The New Year's speeches can be viewed as a very homogeneous genre because of their a) aim, b) form, and c) tradition. As for the aim (a), the goal of the speech is usually to summarize main events of the past year, mention perspectives of a near future, and express best wishes to the inhabitants of the state. The speeches have a very steady form (b), they are prepared in advance and read by the president. The tradition (c) of this kind of speeches is very long in Czech Republic (former Czechoslovakia) – it has started since 1949 and continued up to now (except of 1993 when no president was in office). Obviously, the strong homogeneity of the genre facilitates the authorship analysis because a host of boundary conditions is eliminated.

One can expect two contradictory "powers" which should have the greatest impact on the frequency structure of presidential speeches. On the one hand, the official and ceremonial character of this event should lead to the high uniformity of texts and, consequently, to the high similarity of frequency structures. On the other hand, presidents are usually persons with a strong individuality; as politicians, they have to be able to express their uniqueness and specificity, so, one can expect great differences among them.

For the measurement of frequency characteristics two methods were used: 1) the lambda measurement proposed by Popescu et al. (2010, 2011) and 2) the vocabulary richness index *R1* (Popescu et al. 2009); both methods are presented in the next section.

2. Methodology

The lambda-indicator expresses one aspect of frequency structure of text. In short, it takes into account both the frequency of words and the relationships among individual frequencies. It can be viewed as an indicator of frequency *technique* used by an author. One of the biggest advantage of this measurement is

the independence of lambda-indicator on the text length (cf. Popescu et al. 2011, pp. 10-12). It is defined as

$$\Lambda = \frac{L(\log_{10} N)}{N} \tag{1}$$

where L is the arc length between the ranked frequencies defined as

$$L = \sum_{r=1}^{V-1} [(f_r - f_{r+1})^2 + 1]^{1/2} \tag{2}$$

where N is the text size (in tokens), f_r is the frequency at rank r and V is the highest rank. The variance of Λ is a complex formula and it is presented in detail in Popescu et al. (2010, 2011).

The vocabulary richness index R_1 is defined as

$$R_1 = 1 - \left(F(h) - \frac{h^2}{2N} \right) \tag{3}$$

where h is the h-point (Popescu, Altmann 2006) and $F(h)$ is the cumulative relative frequency up to the h-point. H-point is defined as

$$h = \begin{cases} r, & \text{if there is an } r = f(r) \\ \dfrac{f(i)r_j - f(j)r_i}{r_j - r_i + f(i) - f(j)}, & \text{if there is no } r = f(r) \end{cases} \tag{4}$$

i.e. that point at which $r = f(r)$, or, if there is no such point, it is computed by means of the second part of formula (4).

The variance of R_1 are computed as follows

$$Var(R_1) = \frac{F(h)\left[1 - F(h)\right]}{N}. \tag{5}$$

Let us illustrate the procedure of comparison of authors in the case of three presidents, namely Gottwald, Novotný, and Klaus, by using lambda-indicator. First, from Table 7 in Appendix both the mean lambdas and variances of means are computed, see Table 1.

Table 1.
Mean lambdas and variances of lambdas of New Year's presidential speeches. Presidents are ordered according to the magnitude of mean lambda.

President	**year**	**n**	**mean(Λ)**	**s^2(Λ)**	**s^2 (Λ)/n**
Klaus	2004-2011	8	1.9292	0.003834	0.000479
Husák	1975-1989	15	1.9211	0.008357	0.000557
Havel	1990-2003	13	1.8882	0.004031	0.000310
Zápotocký	1954-1957	4	1.8818	0.001945	0.000486
Svoboda	1968-1974	6	1.8769	0.005683	0.000026
Gottwald	1979-1953	5	1.8714	0.005268	0.001054
Novotný	1958-1968	11	1.7564	0.004349	0.000395

As can be seen in Table 1, Klaus has the highest mean lambda, while Gottwald and Novotný obtain the lowest lambda values. The first task is to observe, whether the differences of mean lambdas are significant. Because the texts of the same genre in the same language are analyzed, we use the asymptotic *u*-test

$$u = \frac{\Lambda_1 - \Lambda_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad . \tag{6}$$

Specifically, for the comparison of Gottwald and Novotný we obtain

$$u = \frac{1.8714 - 1.7564}{\sqrt{0.001054 + 0.000395}} = 3.02$$

which expresses a significant difference. So, we can state that the frequency structure, expressed by mean lambda, of Gottwlad's speeches is significantly different from those of Novotný. Analogously, if we compare Klaus and Novotný, we obtain $u = 5.85$ which is significant too, however, for Gottwald and Klaus we obtain a non-significant $u = 1.48$. The results reveal that Novotný's speeches have significantly different frequency structure in comparison with both Gottwald and Klaus, while the frequency structures of Gottwald's and Klaus' speeches express similarities.

3. The lambda structure of presidential speeches

Following the procedure presented in the previous section, we obtain the results presented in Table 2.

Table 2
Comparison of mean lambdas in New Year's presidential speeches of Czech or Czechoslovak Presidents (two-sided *u*-test). Bold values express significant differences (significance level $u \leq 1.96$).

President	**Gottwald**	**Zápotocký**	**Novotný**	**Svoboda**	**Husák**	**Havel**	**Klaus**
$\overline{\Lambda}$	1.8714	1.8818	1.7564	1.8769	1.9211	1.8882	1.929
$s^2(\overline{\Lambda})$	0.001054	0.000486	0.000395	0.000026	0.000557	0.000310	0.000479
Gottwald	x						
Zápotocký	0.27	x					
Novotný	**3.02**	**4.22**	x				
Svoboda	0.17	0.22	**5.87**	x			
Husák	1.24	1.22	**5.34**	1.83	x		
Havel	0.45	0.23	**4.96**	0.62	1.12	x	
Klaus	1.48	1.53	**5.85**	**2.33**	0.25	1.46	x

For the sake of better lucidity, the relationships among the presidents can be expressed graphically. Figure 1 represents a small network based on Table 2 in which two presidents are connected, if there is non-significant difference between their mean lambdas (i.e., $u \leq |1.96|$). Presidents with the same number of similarities are put at the same level – for Havel, Gottwald, Husák, and Zápotocký, each obtains five similarities, Klaus and Svoboda four, and Novotný has zero.

At a first sight, the extraordinary position of Novotný is evident – there are no similarities between Novotný's mean lambda value and any other president. Further, Klaus and Svoboda can be seen as counterparts because their frequency structures differ significantly, while they both are connected to the same other presidents. Havel, Husák, Zápotocký, and Gottwald represents the most uniform cluster of this genre with regard to lambda.

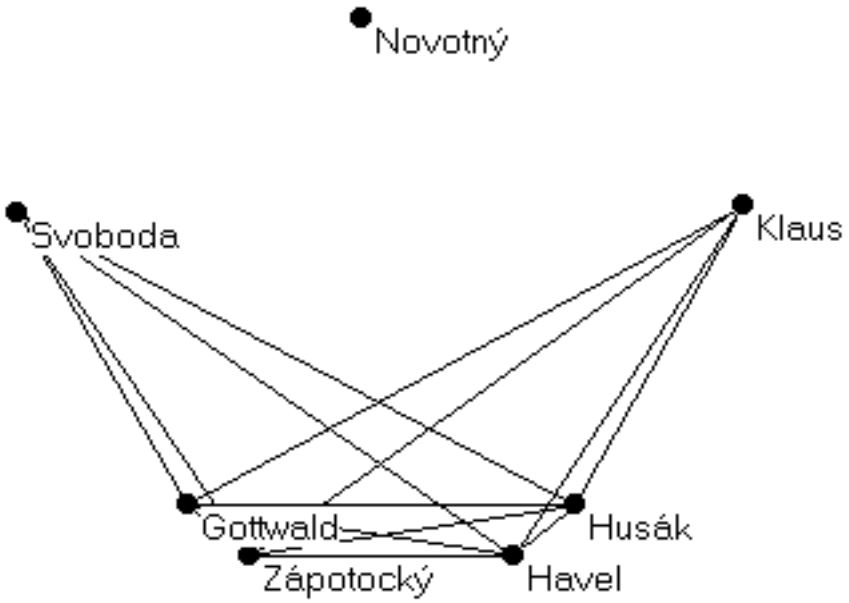


Figure 1 The network in which two presidents are connected, if there is a non-significant difference between their mean lambdas (i.e., $u \leq 1.96$)

For more detailed comparison it is possible to measure a weighted u_w differences among presidents

$$(7) \quad u_{wi} = \frac{\sum u_i}{\sqrt{k}}$$

where k is a number of comparisons. The results based on the formula (7) are presented in Table 3.

Again, the extraordinary position of Novotný is even more obvious. A comparison of weighted differences u_w and mean lambdas reveals that Novotný's position is given by the simplest frequency structure of his speeches, as is illustrated in Figure 2. The values of u_w of the other presidents are located within a relatively small interval <2.70, 5.26> which indicates high homogeneity of this genre with regard the frequency structure expressed by lambda-indicator.

Table 3
The weighted differences u_w of presidents

President	Λ	u_W
Gottwald	1.8714	2.70
Zápotocký	1.8818	3.13
Havel	1.8882	3.61
Husák	1.9211	4.49
Svoboda	1.8769	4.52
Klaus	1.929	5.26
Novotný	1.7564	11.95

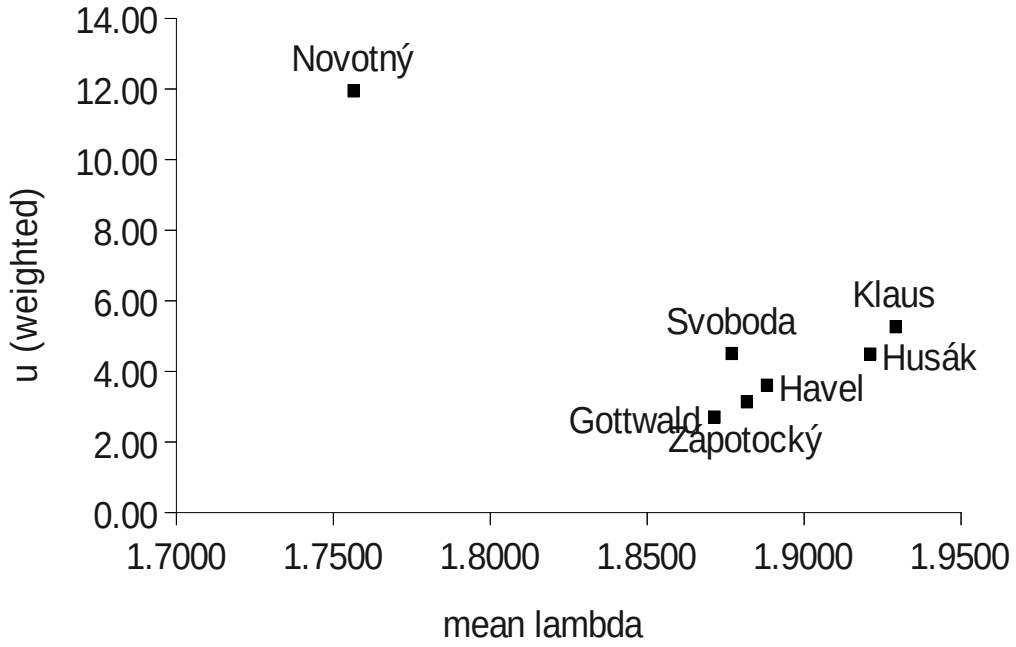


Figure 2. The weighted differences u_w of presidents

4. The vocabulary richness R_1

The computation of vocabulary richness R_1 reveals different ranking of presidents, and very small differences among them – all mean values of R_1 are in the interval <0.8546, 0.8770>, see Table 4.

Table 4
Mean vocabulary richness R_1
and variances of R_1 of New Year's presidential speeches.
Presidents are ordered according to the magnitude of mean R_1.

President	**year**	**n**	**mean(R_1)**	**$s^2(R_1)$**	**$s^2(R_1)$/n**
Svoboda	1969-1974	6	0.8770	0.000556	0.000094
Klaus	2004-2011	8	0.8727	0.000125	0.000016
Husák	1975-1989	15	0.8724	0.000207	0.000014
Havel	1990-2003	13	0.8607	0.000311	0.000024
Zápotocký	1954-1957	4	0.8555	0.000052	0.000013
Novotný	1958-1968	11	0.8552	0.000079	0.000007
Gottwald	1949-1953	5	0.8546	0.000548	0.000110

Performing *u*-tests among all presidents we obtain the results presented in Table 5 and graphically expressed differences in Figure 3.

Table 5
Comparison of vocabulary richness R_1 in New Year's presidential speeches
of Czech or Czechoslovak Presidents (two-sided *u*-test).
Bold values express significant differences (significance level $u = |1.96|$)

President	**Gottwald**	**Zápotocký**	**Novotný**	**Svoboda**	**Husák**	**Havel**	**Klaus**
$\overline{R_1}$	0.8546	0.8555	0.8552	0.877	0.8724	0.8607	0.8727
$s^2(\overline{R_1})$	0.00011	0.000013	0.000007	0.000094	0.000014	0.000024	0.000016
Gottwald	x						
Zápotocký	0.08	x					
Novotný	0.06	0.07	x				
Svoboda	1.57	**2.08**	**2.17**	x			
Husák	1.60	**3.25**	**3.75**	0.44	x		
Havel	0.53	0.85	0.99	1.50	1.90	x	
Klaus	1.61	**3.19**	**3.65**	0.41	0.05	1.90	x

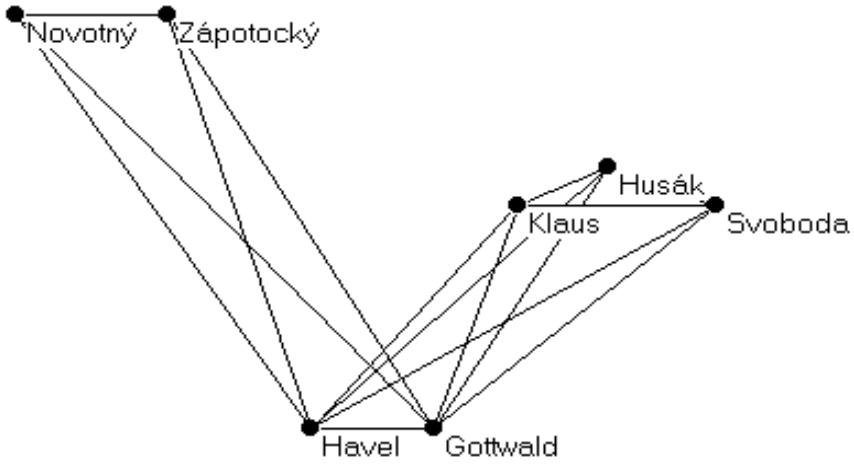


Figure 3. The network in which two presidents are connected,
if there is non-significant difference between their mean R_1 (i.e., $u \leq |1.96)$

Analogously to Figure 1, presidents with an equal number of links are put at the same level and, further, presidents who connect the same other presidents are clustered. As is seen in Figure 3, Novotný is again the president with the lowest similarities (with Zápotocký). Contrary to lambda-measurement, Klaus and Svoboda have non-significant difference of R_1, so, their speeches differ because of frequency technique they used. Finally, Havel and Gottwald are connected to all presidents which means that they are the most conformal authors with regard to vocabulary richness (the author's conformity is analysed in more detail in Section 5)

The computation of weighted u_w differences of R_1 reveals very small differences among presidents, cf. Table 6 and Figure 4.

Table 6.
The weighted differences u_w of presidents

President	R_1	u_w
Gottwald	0.8546	2.22
Havel	0.8607	3.13
Svoboda	0.8770	3.34
Zápotocký	0.8555	3.89
Novotný	0.8552	4.36
Klaus	0.8727	4.41
Husák	0.8724	4.49

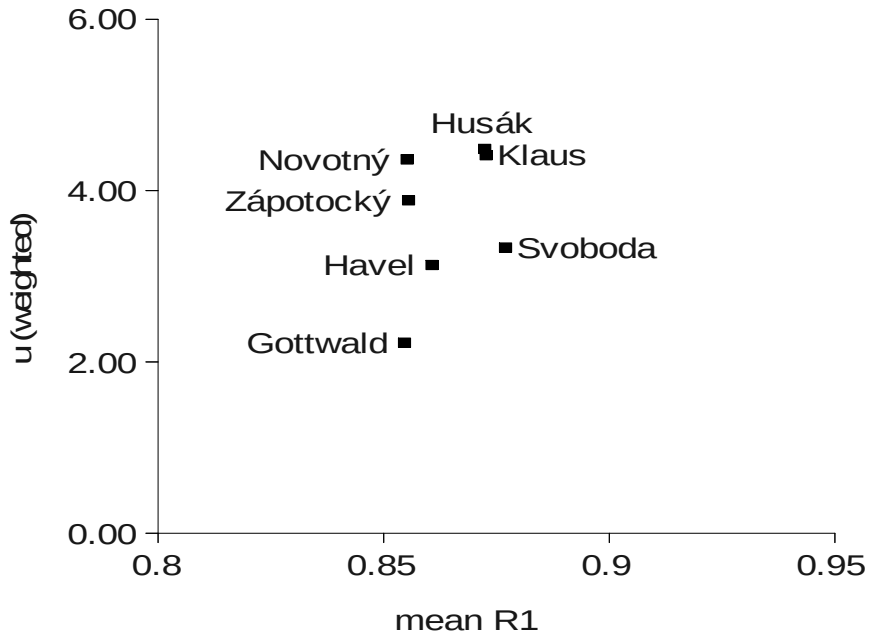


Figure 4. The weighted differences u_w of presidents

The small interval in which all weighted differences u_w lie reflects a high similarity of vocabulary richness. So, the authorship's differences of presidents are caused mainly by the different frequency techniques (i.e. expressed by lambda) which particular presidents used, as is clearly seen in Figure 5.

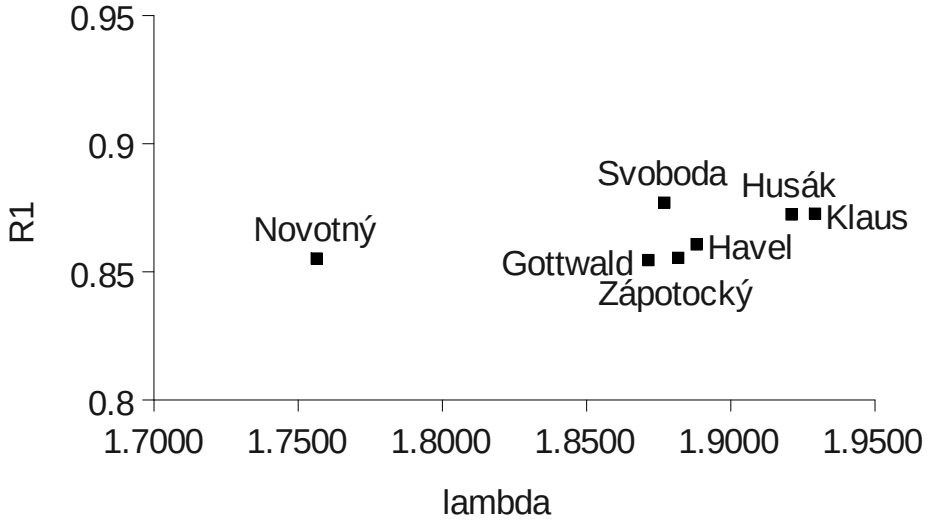


Figure 5. Comparison of lambda-values and R_1.

In Figure 5, all presidents lay in almost horizontal line which means high similarities of R_1. The authorship differences are caused by dissimilar frequency techniques which is indicated by lambda-differences.

5. The measurement of author's conformity within the genre

The properties of graphs (see Figure 1 and 3) enable to propose a *conformity indicator*. It is given by the relative degree of node representing the author, i.e. by the relative number of its links

$$(7) \quad d_{i\,(rel)} = \frac{\sum l_i}{l_{i\,(max)}}$$

where l_i is an observed number of links of the node and $l_{i(max)}$ is the maximum number of links which can the node obtain

$$(8) \quad l_{i\,(max)} = (n-1)x$$

where n is a number of nodes in the network and x is the number of particular networks used for measurement. For illustration, for Havel (based on graphs in Figure 5 and 3, i.e. $x = 2$) we obtain

$$d_{Havel\,(rel)} = \frac{11}{(7-1)2} = 0.92\,.$$

The conformity indicator of all presidents is shown in Table 6

Table 6
Conformity indicator of Czech presidents, based on lambda-measurement and vocabulary richness R_1, expressed by the relative degree of node d_{rel}. The lower d_{rel} , the more original is an author and vice versa.

Presidents	$\boldsymbol{d_{rel}}$
Gottwald, Havel	0.92
Husák	0.75
Klaus, Svoboda, Zápotocký	0.67
Novotný	0.25

Novotný is evidently the most original author among presidents, with regard to both lambda-structure and vocabulary richness. On the other hand, Gottwald and Havel are the most conformal ones. As for Havel, this result is a little surprise – one could expect that Havel, as the world-famous dramatist, should strive for the greatest language originality. However, if Havel's frequency technique appears to be conformal also in the other genres, it should mean that frequency conformability can be taken as a characteristic feature of his language usage. Of course, the conformability or frequency technique itself has nothing to do with a content and literary quality of his texts. Contrariwise, the same frequency technique can be used for extremely different purposes, as our results clearly manifest – it is striking that the most conformal authors (i.e., Gottwald, Havel) are persons which can be viewed as political and personal counterparts: Gottwald was a professional politician, leader of communist coup, dictator, while Havel has been a writer, long-term leader of democratic opposition in communist Czechoslovakia, democrat, humanist.

6. Conclusion

The analysis of lambda-structure and vocabulary richness of presidential speeches reveals surprisingly high number of similarities among presidents. This indicates that the tendency to uniformity prevails the need to express individuality of particular persons. Moreover, both measurements do not unveil the impact of period (the speeches do not reflect the changes in sixty years of language development) or political regime (Gottwald, Zápotocký, Novotný, Svobooda, and Husák are representatives of communist totality, while Havel and Klaus represent democracy).

Acknowledgement
I thank Jaroslav David for providing me the data which have been used for the analysis. This work has been also supported by the Czech Science Foundation, grant no. P406/11/0268.

References

Popescu, I.-I., Altmann, G. (2006). Some aspects of word frequencies. *Glottometrics 13, 23-46.*

Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B.D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., Vidya, M.N. (2009). *Word frequency studies.* Berlin-New York: Mouton de Gruyter.

Popescu, I.-I., Mačutek, J., Altmann, G. (2010). Word forms, style and typology. *Glottotheory 3(1), 89-96*

Popescu, I.-I., Čech, R., Altmann, G. (2011). *The lambda-structure of texts.* Lüdenscheid: RAM.

Appendix

Table 7
New Year's presidential speeches.

President	year	N	V	h	L	Λ	var(Λ)	R1	var(R1)
Gottwald	1949	1413	828	10	881.4247	1.9650	0.000395	0.8882	0.000089
Gottwald	1950	2205	1115	15	1188.516	1.8021	0.000377	0.8342	0.000077
Gottwald	1951	2211	1121	13.5	1186.033	1.7941	0.000352	0.8359	0.000074
Gottwald	1952	1817	971	11.67	1053.653	1.8901	0.000240	0.8454	0.000085
Gottwald	1953	1651	911	11.5	977.842	1.9058	0.000399	0.8692	0.000086
Zápotocký	1954	2590	1272	14	1387.437	1.8285	0.000316	0.8502	0.000059
Zápotocký	1955	1570	846	9	931.7357	1.8966	0.000200	0.8627	0.000087
Zápotocký	1956	2904	1446	14.5	1567.306	1.8690	0.000153	0.8485	0.000053
Zápotocký	1957	2486	1331	13	1415.178	1.9329	0.000219	0.8606	0.000058
Novotný	1958	1592	863	10.5	925.0773	1.8606	0.000391	0.8701	0.000086
Novotný	1959	2119	1066	14.25	1145.578	1.7982	0.000231	0.8563	0.000073
Novotný	1960	2755	1300	14.67	1409.963	1.7606	0.000234	0.8550	0.000055
Novotný	1961	1583	876	11.5	920.397	1.8603	0.000457	0.8630	0.000093
Novotný	1962	2709	1289	14	1374.971	1.7423	0.000276	0.8520	0.000056
Novotný	1963	1954	973	13	1033.525	1.7407	0.000404	0.8570	0.000078
Novotný	1964	2915	1338	17	1399.033	1.6628	0.000304	0.8472	0.000055
Novotný	1965	2290	1092	14.5	1162.03	1.7049	0.000456	0.8664	0.000064
Novotný	1966	3263	1494	17	1575.538	1.6965	0.000209	0.8390	0.000050
Novotný	1967	2572	1210	15	1281.981	1.6998	0.000212	0.8513	0.000060
Novotný	1968	2293	1157	14.5	1224.202	1.7941	0.000195	0.8505	0.000069
Svoboda	1969	2059	1081	13.67	1146.529	1.8452	0.000317	0.8462	0.000078
Svoboda	1970	2186	1097	14.5	1171.538	1.7898	0.000276	0.8505	0.000073
Svoboda	1971	1554	883	11	929.011	1.9079	0.000407	0.8755	0.000088
Svoboda	1972	454	295	6	310.0376	1.8145	0.000963	0.8965	0.000270
Svoboda	1973	508	342	6	358.4382	1.9092	0.000432	0.8937	0.000239
Svoboda	1974	429	311	6	325.1071	1.9949	0.000639	0.8998	0.000284
Husák	1975	1520	789	10.5	845.9319	1.7708	0.000248	0.8613	0.000095
Husák	1976	1486	815	10.67	884.6006	1.8883	0.000612	0.8573	0.000100
Husák	1977	1281	682	10	730.9015	1.7731	0.000514	0.8735	0.000108
Husák	1978	1598	825	13.25	897.1975	1.7986	0.000464	0.8509	0.000102
Husák	1979	1322	778	9.667	836.7562	1.9756	0.000346	0.8757	0.000101
Husák	1980	1377	803	9.5	871.703	1.9871	0.000464	0.8636	0.000102
Husák	1981	1548	863	11	924.6461	1.9053	0.000370	0.8647	0.000093
Husák	1982	1155	671	9	716.5085	1.8999	0.000440	0.8818	0.000112

Husák	1983	1128	661	10	705.1472	1.9081	0.000366	0.8715	0.000127
Husák	1984	1032	661	8.5	715.8022	2.0903	0.000514	0.8974	0.000115
Husák	1985	1378	795	11	851.3336	1.9394	0.000720	0.8661	0.000106
Husák	1986	1288	756	10	813.2448	1.9636	0.000664	0.8672	0.000110
Husák	1987	1491	839	11.33	895.8157	1.9067	0.000321	0.8714	0.000095
Husák	1988	776	518	6.5	549.2147	2.0453	0.000536	0.9061	0.000137
Husák	1989	860	536	7	575.5355	1.9638	0.000245	0.8773	0.000149
Havel	1990	2347	1225	15	1287.602	1.8491	0.000299	0.8673	0.000063
Havel	1991	2421	1241	14.33	1323.707	1.8502	0.000343	0.8516	0.000064
Havel	1992	3278	1638	15	1774.774	1.9034	0.000138	0.8437	0.000047
Havel	1994	2746	1353	16	1428.073	1.7883	0.000339	0.8518	0.000057
Havel	1995	3240	1594	16	1724.62	1.8686	0.000187	0.8407	0.000049
Havel	1996	2750	1392	15	1471.607	1.8405	0.000255	0.8391	0.000059
Havel	1997	598	397	7	419.7517	1.9490	0.000503	0.9055	0.000196
Havel	1998	1312	720	12	754.9176	1.7940	0.000579	0.8643	0.000118
Havel	1999	1723	1012	11	1067.827	2.0057	0.000249	0.8738	0.000079
Havel	2000	2021	1111	13	1172.44	1.9177	0.000171	0.8671	0.000071
Havel	2001	1587	893	12.5	946.0883	1.9080	0.000500	0.8526	0.000100
Havel	2002	1926	1062	12	1126.907	1.9219	0.000293	0.8629	0.000075
Havel	2003	1941	1089	13	1150.816	1.9495	0.000294	0.8689	0.000074
Klaus	2004	913	527	10	560.2908	1.8168	0.000616	0.8642	0.000169
Klaus	2005	979	604	9.5	637.3513	1.9471	0.001002	0.8714	0.000147
Klaus	2006	845	526	9.5	556.1016	1.9262	0.000593	0.8676	0.000179
Klaus	2007	799	534	9	555.8242	2.0192	0.000632	0.8867	0.000172
Klaus	2008	914	559	8.75	584.8158	1.8945	0.000675	0.8635	0.000160
Klaus	2009	870	551	9	574.851	1.9423	0.000667	0.8914	0.000151
Klaus	2010	910	557	8.67	584.0399	1.8991	0.000564	0.8611	0.000162
Klaus	2011	891	568	8	600.5498	1.9900	0.000412	0.8754	0.000151

Kelih, E. et al. (eds.), Issues in Quantitative Linguistics, Vol. 2, 88-93

A stylostatistical study using the sentence length criterion

Atul S. Inamdar
S. G. Prabhu–Ajgaonkar

Introduction

Stylistics and the statistical study of text have become crossroads for the interest of literary scholars, critics, linguists, psychologists, sociologists, mathematicians and statisticians. The mathematician or statistician finds in the study of text a body of material to use for testing and refining distributional and discriminatory formulae. The Sociologist is interested in the text as it reflects social factors and developments in the history of ideas. The psychologist investigates the text as a product of the highest and most distinctive capacity of the human mind. The linguist studies the text structure in order to discover patterns and variability both in the language system and the use of language. Finally, the literary scholar and critic find in the text the focus of their professional interest in artistic language. This startling variety of approaches to the text and its style leads to widely disparate results.

The statistical investigation of text and the text styles can serve all the specialists mentioned above. However**,** it is most directly concerned with the description and explanation of features inherent in the text, their organization and variability. Though a scholar may be highly competent in his own area, he is often very much an amateur in one or more of the disciplines that find their meeting ground in statistical stylistics, or stylo-statistics.

Sentence length

A sentence means a meaningful arrangement of different words. An author is describing an event or thought or making conversation with the reader using different types of sentences. Thus a sentence could be defined as a set of words that is complete in itself, conveying a statement, question, exclamation, or command and typically containing a subject and predicate.

Sentence length depends on the number of words in the sentence. However, some researchers use the number of clauses as the yardstick of sentence length.

Generally Writers use sentences of different lengths. The author's imagination, writing skill, knowledge, and memorization affect the sentences used by him.

A review of past studies

Udney Yule (1939) attacked the problem of authorship from the angle of variation in sentence length, and this appeared to be a much more fertile method than the other methods to study style of author. He showed that a frequency distribution of sentence length (i.e. number of words between successive full stops) is of skew type. Williams (1939) and Herdan (1964) showed that the sentence length reflects the style of author. Sentence length is also the well-known factor of readability due to Flesch (1948).

Williams (1939) and Wake (1957) indicated that the distribution of sentence length in terms of words follows a log-normal distribution. For this study they used the social books of three different authors: Wells, Shaw and Chesterton. Rao Subba (1960) studied sentence length in eight works in Kannada Prose of three different authors. It is noted that authors displayed significant differences among their writings.

It is pertinent to mention the contributions of Altmann (1988), Best (2001, 2001a,b, 2002, 2003), Busch (2002), Kelih, Grzybek (2004), Levickij, Pavlyčko, Semenyuk (2001), Niehaus (1997, 2001), Roukk (2001), Williams (1939), Wittek (2001) in this field .

Sentence length distribution is, of course, not the only characteristic of writer's style. One can scrutinize the sequence of sentence lengths, the grammatical complexity (dependence structure) of sentences, the readability of sentences, their referential structure, the valence of some words in the sentence, etc. However, in this study we shall restrict ourselves to the distribution of sentence lengths.

About the author

V. S. Khandekar was a famous Indian Marathi novelist. He wrote nearly 75 books. His main aim was to give happiness to the society through literature and art and to remove inequality among the people.

He was awarded the "Gohar" gold medal for his picture story "Chhaya" in 1936. The Indian Government awarded him the prestigious "Dnyanpeeth" award for his novel "Yayati" in the year 1974. In ancient Indian literature, especially mythology, there are numerous stories which throw light on different tendencies of individuals and society and at the same time the consequences of these tendencies. Of

these stories, “Yayati” is well known. Taking inspiration from this story V.S. Khandekar wrote the novel “Yayati.” This novel can be said to be the magnum opus of V.S. Khandekar.

Data

Here we are studying the style of the author using the “sentence length criterion” in his selected novel by applying Smirnoff’s test.

The novel “Yayati” has been divided into two distinct parts and from each part a sample of 2000 sentences has been selected. The first part consists of beginning pages of the novel and the second part comprises end pages of the novel. The number of words in each sentence was calculated, and the frequency f_x of sentences containing x words was determined. This is shown in Table 1 below.

Table 1
Cumulative frequencies of sentence length: Novel Yayati:
Sample I and II

Sr.No.	Class Interval	Cumulative Frequency F_1(Front Side)	F_1/n_1	Cumulative Frequency F_2(Back Side)	F_2/n_2	$\lvert F_1/n_1 - F_2/n_2 \rvert$
1	0.5 to 5.5	610	0.3050	713	0.3565	***0.05150***
2	5.5 to 10.5	1500	0.7500	1558	0.7790	0.02900
3	10.5 to 15.5	1821	0.9105	1862	0.9310	0.02050
4	15.5 to 20.5	1940	0.9700	1958	0.9790	0.00900
5	20.5 to 25.5	1979	0.9895	1985	0.9925	0.00300
6	25.5 to 30.5	1991	0.9955	1992	0.9960	0.00050
7	30.5 to 35.5	1997	0.9985	1996	0.9980	0.00050
8	35.5 to 40.5	1998	0.9990	1998	0.9990	0.00000
9	40.5 to 45.5	1999	0.9995	1999	0.9995	0.00000
10	45.5 to 50.5	2000	1.0000	2000	1.0000	0.00000

Methodology and procedure

Smirnoff (1939) proposed a test criterion which can be applied independent of the type of distribution. The problem is to investigate whether the two samples

$x_1 \leq x_2 \leq ... \leq x_{n1}$
$y_1 \leq y_2 \leq ... \leq y_{n2}$

are drawn from the same population or not. For this purpose we define two staircase lines $S(x)$ given by,

$$S(x) = \begin{cases} 0 & for\ x < x_1 \\ k/n & for\ x_k \leq x \leq x_{k+1} \quad for\ k = 2,3,...,n_1 - 1 \\ 1 & for\ x_{n_1} \leq x \end{cases}$$

and a similar one for $S(y)$.

If we introduce the quantity

$D(n_1, n_2) = \lim_{-\infty < x,y < \infty} \text{Sup} \left| S(x) - S(y) \right|$.

When the samples belong to the same category, then

$\text{Prob}\{D(n_1,n_2) \leq \boldsymbol{\lambda\sqrt{1/n_1+n_2}}\} \rightarrow \boldsymbol{\Phi(\lambda)}$ **,**

$$\Phi(\lambda) = \sum_{k=-\infty}^{\infty} (-1)^k \exp(-2k^2\lambda^2)$$

where n_1 and $n_2 \rightarrow \infty$ such that n_1/n_2 remains constant. This result is true whatever the distribution function may be.

Now $D(n_1,n_2)$ can be written as $\hat{D} = \text{Max} \left| (F_1/n_1) - (F_2/n_2) \right|$, F_1 and F_2 are the cumulative frequencies and n_1 and n_2 are the sample sizes. The differences $(F_1/n_1) - (F_2/n_2)$ are calculated at regular intervals. The maximum of the absolute difference furnishes the test statistic $\hat{D}$. The critical values of D can be approximated for a medium to a large sample of sizes $(n_1 + n_2) > 35$ by $D_{(\lambda)} = \boldsymbol{\lambda\sqrt{(1/n_1+1/n_2)}}$ where λ is a constant depending on the level of significance. If the value of $\hat{D}$ determined from the two sample groups equals or exceeds the critical value of $D_{(\lambda)}$ then a significant difference exists between the distribution of the two populations.

From Table 1, it is noted that the calculated $\hat{D}$ = 0.05150 which is the maximum of above differences. The critical value of $D_{(\lambda)}$ is = 0.358 at 5% level of significance.

Conclusion

In the present paper it has been noted that the two samples are drawn from the same population. In other words, the author's style remains same throughout the novel as regards the "sentence length" criterion tested by Smirnoff's test.

References

Altmann G. (1988). Verteilungen der Satzlängen. In: *Glottometrika 9, 147-169*. (Ed. Klaus-Peter Schulz). Bochum: Brockmeyer.

Best, K.-H. (ed.) (2001). *Häufigkeitsverteilungen in Texten*. Göttingen: Peust & Gutschmidt.

Best, K.-H. (2001a). Probability distributions of language entities. *Journal of Quantitative Linguistics 8, 1-11*.

Best, K.-H. (2001b). Wie viele Wörter enthalten Sätze im Deutschen? Ein Beitrag zu den Sherman-Altmann-Gesetzen. In: Best 2001, 167-201.

Best, K.-H. (2002). Satzlängen im Deutschen. Verteilungen, Mittelwerte, Sprachwandel. *Göttinger Beitrage zur Sprachwissenschaft 7, 7-31*.

Best, K.-H. (2003). *Quantitative Linguistik. Eine Annäherung.* 2. überarbeitete and erweiterte Auflage. Göttingen: Peust & Gutschmidt.

Busch, A. (2002). *Zur Entwicklung der Satzlänge in deutscher Fachsprache.* Staatsexamensarbeit, Göttingen.

Flesh R. (1948). A new readability yardstick. *Journal of Applied Psychology 32, 221-233*

Herdan, G. (1964). *Quantitative Linguistics*. London: Butterworth

Kelih, E., Grzybek, P. (2004). Häufigkeiten von Satzlängen: Zum Faktor der Intervallgröße als Einflussvariable (am Beispiel slowenischer Texte). *Glottometrics 8, 23-41*.

Köhler, R. (1995). *Bibliography of Quantitative Linguistics*. Amsterdam-Philadelphia: Benjamins.

Köhler, R., Altmann, G., Piotrowski, R.G. (eds.) (2005). *Quantitative Linguistics. An International Handbook.* Berlin-New York: de Gruyter.

Levickij, V.V., Pavlyčko, O.O., Semenyuk, T.G. (2001). Sentence length and sentence structure as statistical characteristics of style in prose. In: Uhlířová, L.

et al. (eds.), *Text as a linguistic paradigm: levels, consttuents, constructs. Festschrift in Honour of Ludek Hřebíček: 177-185*. Trier: WVT.

Niehaus, B. (1997). Untersuchung zur Satzlängenhaufigkeit im Deutschen. In: *Glottometrika 16, 213-275*. (Ed. K.-H. Best). Trier: Wissenschaftlicher Verlag Trier.

Niehaus, B. (2001). Die Satzlängenverteilung in literarischen Prosatexten der Gegenwart. In: L. Uhlířová, G. Wimmer, G. Altmann and R. Köhler (eds.), *Text as a Linguistic Paradigm: Levels, Constituents, Constructs. Festschrift in Honour of Luděk Hřebíček: 196-214*. Trier: Wissenschaftlicher Verlag Trier.

Rao Subba (1960). A study into the sentence length as a statistical characteristic determining the prose style of an author. *The Half Yearly Journal of the Mysore University, New series,Section A-Arts,Vol XX, No. 1, 1-12.*

Roukk, M. (2001). Satzlängen in Russischen. In: Best 2001, 211-218.

Sachs, L. (1984). *Applied Statistics. A Handbook of Techniques.* New York-Berlin-Heidelberg-Tokyo: Springer.

Sichel, H.S. (1974), On a distribution representing sentence-length in prose. *Journal of the Royal Statistical Society (A) 137, 25-34.*

Smirnoff, N. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bulletin Mathematique de l'Université de Moscou, serie Internationale, II, 1-16.*

Yule, G. U. (1938). On sentence length as a statistical characteristic of style in prose with applications to two cases of disputed authorship. *Biometrika 30, 363-390.*

Wake. C.W. (1957). Sentence length distribution of Greek author. *Royal Statistical Society 120, Series A , 331-346.*

Williams, C.B. (1939). A note on the statistical analysis of sentence-length as a criterion of literary style. *Biometrika 31, 356-361.*

Wittek, M. (2001). Zur Entwicklung der Satzlänge im gegenwärtigen Deutschen. In: Best 2001, 219-247.

Kelih, E. et al. (eds.), Issues in Quantitative Linguistics, Vol 2, 94-103

Determining phonetic symbolism of the text

Nadija L. Lvova

The problem of phonetic symbolism goes back to the works of classical philosophers, who were the first to notice that speech sounds recall certain steady and common associations for all speakers of the same language. The interest in this issue has not diminished over the centuries, but the scientific development of sound symbolic matters became possible only after the introduction and application of objective psycholinguistic methods. In the 20th century European and American science had different views on phonetic symbolism, in particular on its nature and methods of investigation. In Europe phonetic symbolism studies were characterized by a great number of works dedicated to the theoretical issues of sound symbolics (Sieberer 1974; Kronasser 1952; Ullmann 1959; 1964). In the USA and Canada much attention was paid to the experimental studies of phonetic symbolism, especially to working out of the methods of experiments (Tsuru, Fries 1933; Taylor, Taylor 1965; Brown, Nuttal 1959). Further studies of phonetic symbolism led to the corroboration of phonosemantic properties of sounds in various languages both within one language family and in different language families. Later, scholars applied phonosemantic analysis for creating the names of some goods (Hinton, Nichols, Ohala 1994) and for working out the names of trademarks. Thus, the studies of interrelation between sound and meaning still go on. In this article we put forward the hypothesis about the existence of phonetic (= submorphemic) meaning and we make an attempt to demonstrate the interdependence between the usage of certain sounds and the motif of the text.

Two basic types of meaning are commonly differentiated in linguistics – lexical and grammatical – which correspond to lexical, morphological and syntactical levels of the language. The units of the lower phonetic level were considered to be meaningless. But with the beginning of experimental studies of sound symbolism the need to single out one more type of meaning emerged – the phonetic one. If the meaningfulness of a language unit is understood as a symbolic meaning, then one may admit that sound symbolism in the language is understood as the significance of the phonetic form, i.e. *phonetic symbolism*.

O. Zhuravliov, a Russian linguist, assumed that if individual sounds possessed symbolic meaning, then their combinations were sure to have it, too (Zhuravliov 1974: 117). Phonetic symbolism (PhS) is regarded as a pooled estimate of symbolic sounds which are constituent parts of word sounding. Each of these sounds has its own phonetic symbolism that can be measured with the help of the "semantic differential" method using different categories. Zhuravliov proposed a special formula which takes into account different positions or roles in a word (initial sounds, stressed sounds and repeated ones, those possessing peculiar

psychological weight) and determined the PhS of several thousands Russian words using an electronic computer.

V. Levickij, an outstanding Ukrainian linguist, objected to some of the principles laid down by Zhuravliov. First of all, the approach in the selection of words for statistic processing is subjective. It was assumed that strongly marked phonetic meaning was likely to be found in those lexemes and that is why all the words under analysis possessed PhS. But it was not clear to what extent and in what layers of the vocabulary phonetic symbolism was inherent. The interpretation of the correspondence between the PhS and significative meaning of the word was quite subjective, too (Levickij 1998: 84-85). Thus, the methods of determination and interpretation of phonetic meaning require further investigation.

If phonetic symbolism exists, it has to manifest itself in speech. Since phonetic symbolism is the meaningfulness of sound form, manifestation of sound symbolism is supposed to be found more easily in poetic language where the organization of sound form acquires a particular importance. Such texts were analyzed in Zhuravliov's works. But there was no clear evidence of links between the phonetic symbolism of a poetic text and its emotive and semantic side. It may be explained by the subjective approach to the solution of this problem. It is also worth mentioning the attempt made by A. Shtern to find the connection between the connotative meaning of a poetic text and its overall phonetic meaning. She offered a clear-cut method of determining this connection that is based on the deviation of sounds' frequency in a verse from their normal speech frequencies (Shtern 1969). Keeping on the same course of research, Zhuravliov tried to find a way to measure the phonetic symbolism of a poetic text and to compare the results of such studies with the total emotive and semantic content of the lyrics. The basis of the analytic procedure was the comparison of the average estimates of sound symbolism with the sound frequencies' deviation from the norm. But the author himself admitted that the suggested analysis procedure was "too rough" (Zhuravliov 1974: 113) and had some drawbacks. One cannot but notice the element of uncertainty and subjectivity which is observed during the analysis of a literary work. The same poem, for example, may receive different interpretations and its perception may be influenced by time and social factors. The analysis was carried out on texts in which one theme, feeling or mood was brightly expressed or dominated. Widely known works were selected for the analysis so that "the intuitive assessment of their general content caused no doubt" (Zhuravliov 1974: 109). Thus, the intuitive descriptive characteristics of the text content show the subjective approach of the researcher. Besides, the analysis procedure doesn't always consider such basic circumstances of phonetic text organization as sound position in the word, sound repetitions, sound stress etc. But despite these drawbacks, the method of determining the phonetic meaning of a poetic text in the Russian language introduced by Zhuravliov is a certain contribution to the development of the idea of sound symbolism.

Applying Zhuravliov's formula for calculating the phonetic meaning of German words, Levickij suggested an objective procedure for interpreting the received data on the PhS of these words. Having determined the average value of the phonetic symbolism, all cases exceeding the average were considered to be the signs of a certain feature whereas all the other cases showed the lack of this feature. In this way it was shown that 50,5% of the 617 words under study (Kushneryk, Levickij 1986) (according to at least one category of Osgood differential) are characterized by significant and strongly marked phonetic symbolism. Besides, "word sounding may be comprised of such sounds which, according to the accepted statistical procedure, will all together provide the word with some phonetic symbolism" (Levickij 1998: 87).

The idea of ascribing separate sounds and consonant clusters certain symbolic meanings is still supported and developed by the research of other contemporary linguists (Magnus 1999; Abelin 1999; Hinton, Nichols and Ohala 1994; Lowrey, Shrum 2007 etc.) but they do not focus on the correlation that might exist between the phonetic constituents of the text and its motif or meaning.

In view of the abovementioned facts, we decided to work out our own methods of calculating the value of PhS, to verify the relationship between phonetic form and text content, and to study the correspondence between the phonetic structure and phonetic symbolism of English poetry, prose and newspaper fragments.

According to Zhuravliov, phonetic symbolism is based on attributive characteristics. Connotative and attributive components are aspects of the lexical meaning of a word. Phonetic form will correspond with the meaning if a correspondence between the phonetic symbolism and connotative-attributive aspect of lexical meaning can be observed. Since our task is to make the investigation more objective, we propose to use the data derived from statistical calculations in previous studies, namely psycholinguistic analysis and phonosemantic research of three styles in English.

Our previous research dealt with the psycholinguistic study of sound-symbolic properties of 30 English initial consonant clusters (Lvova 2004). The study of these properties was performed with the help of the semantic differential, one of the major methods of experimental psycholinguistics introduced by Osgood and widely used in phonosemantics to investigate the symbolic significance of sounds. The psycholinguistic analysis proved the existence of symbolic meanings with initial consonant clusters according to six categories of Osgood's differential ("power", "evaluation", "activity", "evenness", "rigidness", "size"). As a result of the method of "semantic differential" according to these six categories, average estimates of English initial consonant clusters in these categories became known which correspond to the connotative characteristics of the given combinations (Lvova 2004). Application of the statistical method (chi-square test) provided us with the opportunity not only to state the existence of the interaction between phonetic and semantic components in the English initial consonant

clusters but also to reveal the "meanings" or attributes certain clusters are characterized by. The chi-square test helped to identify any statistically significant phonosemantic relationships. The calculations were conducted by means of the following formula proposed by Levickij:

$$(1) \qquad \chi^2 = \frac{(ad-bc)^2 N}{(a+c)(b+d)(a+b)(c+d)},$$

where *a,b,c* and *d* are the empirical data from the four-field tables used in the statistical analysis and *N* the total number of observations. For purposes of statistical analysis, it is often useful to present the data in the form of the "alternative distribution", that is, in tables consisting of four fields (two columns and two lines). Four-field tables were constructed in cases where the empirical datum of the cluster was greater than its theoretically expected datum and the difference between them was positive. Relationships which turned out to be statistically significant ($\chi^2 \geq 3{,}84$) were evidence for the existence of real correspondences between the semantic categories and the initial consonant combinations. When the empirical datum was less than the theoretically expected value, the null hypothesis was recognized; that is, that there was no relationship between the phenomena under investigation.

Thus, we offer a sample of the four-field table that illustrates the distribution of occurrence in the cluster *Bl-* and other English consonant clusters with the category "power" (Table 1).

Table 1
Frequency distribution of the cluster *Bl-* and other consonant combinations in the category "Power"

	Weak		Other (strong)		Total
Bl-	40	*a*	*b*	29	69
Other clusters	732	*c*	*d*	1180	1912
Total	772			1209	1981

After applying formula (1), we determined the value of the chi-square for the cluster *Bl-* ($\chi^2 = 10{,}85$). These data corroborate the "weakness" of the consonant cluster. In a similar way we determined the correspondence between the semantic categories and 30 initial consonant clusters in English.

The results of the chi-square test also helped to corroborate the assumption that certain initial consonant clusters are used to express "pleasant" and "unpleasant" motifs. Thus we decided to apply the data received in the analysis for our present research where we would focus on determining the correlation between

the use and symbolic meaning of English consonant clusters on the one hand and the motif of English texts belonging to different styles on the other.

Since texts of poetry and prose used in our experiment were grouped according to their content, emotional mood and influence on the reader and were defined as texts with "positive connotation" (PC) and "negative connotation" (NC) (it can be admitted to be the only concession to subjectivity in this experiment), we can assume that these text characteristics coincide with the category "evaluation" of the "semantic differential" ("pleasant" – "unpleasant"). Thus, the aim of the research is as follows: a) To study the correlation between the connotative characteristics of the clusters which are most frequent in the three styles and the content of the texts and b) to determine and compare the phonetic symbolism of the three kinds of texts: poetry, prose and newspaper.

The material of investigation was a corpus consisting of the following text fragments: 12 poetic texts, eight prose fragments and 10 newspaper articles under the heading "National Politics". All the fragments within these three genres are chronologically and thematically similar. So the requirements for homogeneity which are important for sampling in statistics were complied.

Determining phonetic symbolism of poetic texts. The following procedure is proposed to perform the tasks. Poetic texts have been divided into two groups. The group of texts with PC consisted of the lyrics *The Man Against the Sky* by E.A Robinson., *The Summer Holds* by W.H Auden., *Tulips* by S. PLath, *A Winter's Tale* by D.A. Thomas, *The River* by H. Crane, *Baile and Aillinn* by W.B. Yeats The group of fragments with NC included such poems as *Strange Meeting* by W. Owen, *The Waste Land* by T.S. Eliot, *General William Booth Enters into Heaven* by V. Lindsay, *After the Funeral* by D. Thomas, *For the Union Dead* by R. Lowell, *Dead Man's Dump* by I. Rosenberg.

The results of the poetic texts' analysis show a certain connection between the clusters *Fl-* ($\chi^2 = 6{,}93$), *Gl-* ($\chi^2 = 6{,}341$), *Pl-* ($\chi^2 = 10{,}73$) and "pleasant" lyrics and between the clusters *Bl-* ($\chi^2 = 12{,}33$), *Kr-* ($\chi^2 = 5{,}71$), *Fr-* ($\chi^2 = 5{,}89$) and "unpleasant" poetic fragments under consideration. Estimates of χ^2 indicate to what extent empirical frequencies of sound combinations *Fl-*, *Gl-*, *Pl-*, *Bl-*, *Kr-* and *Fr-* differ from their theoretical frequencies. In other words, the sums of χ^2 are objective indicators of the excess of cluster frequencies that occur in the text over the expected ("theoretical") frequencies.

Each of these clusters received its own average rating according to six categories of the "semantic differential" (Lvova 2004: 50-51). Since all the clusters were estimated on the five-point scale (five categories), the point 3 was considered neutral, all the points less than three symbolized the notion "unpleasant" and more than three "pleasant" ("evaluation" category). It is reasonable to apply the centering method (to subtract 3 from each average rating), which gives the estimates on the "negative" pole a minus sign and those with the "positive" meaning a plus sign. The average ratings of the initial consonant clusters according to the category "evaluation" are as follows:

Fl- 0,6	*Bl-* -0,2
Gl- 0,1	*Kr-* -0,6
Pl- 0,5	*Fr-* -0,6

Thus, the clusters *Fl-, Gl-, Pl-* are pleasant, the most frequent and correspond to the poetic texts with positive connotation which is corroborated by the significant correlation between these clusters and "pleasant" poems ($\chi^2 > 3{,}84$). We assume that the phonetic symbolism of the poetic texts with *PC* can be found if the χ^2 value of every cluster is multiplied by its average rating on the "evaluation" category and then the sum total of all the results is calculated. Thus, we offer to use the following formula:

$$(2) \qquad PhS_{\chi 2} = (\chi^2{}_1 \times A_1) + (\chi^2{}_2 \times A_2) + (\chi^2{}_3 \times A_3)\ldots,$$

where χ^2 is the chi-square test value for a certain cluster and A is its centered rating (see the figures above). Consequently, *PhS* for poetic texts with positive semantics equals:

$$PhS_{\chi 2} = (6{,}93 \times 0{,}6) + (6{,}34 \times 0{,}1) + (10{,}73 \times 0{,}5) = 10{,}2.$$

A similar procedure was applied to the poems with *NC*. First, the products of χ^2 values for consonant clusters *Bl-, Kr-* and *Fr* and their average ratings on the "evaluation" category were calculated, then the sum total of the products was found:

$$PhS_{\chi 2} = (12{,}33 \times -0{,}2) + (5{,}71 \times -0{,}6) + (5{,}89 \times -0{,}6) = -9{,}4.$$

The negative result received is the best corroboration of the "unpleasant" nature of poems with *NC*. After comparing the phonetic symbolism of poems it is possible to conclude that PhS of poems with *PC* is larger than *PhS* of poems with *NC*. It is statistically corroborated that connotative characteristics of English initial consonant clusters are interconnected with the connotation of poetic texts.

Determining phonetic symbolism of prose texts. The analysis of prose texts was based on the procedure similar to the poetic texts. The group of texts with PC comprised the following fragments: *A Woman of Substance* by B.T. Bradford.; *The Grapes of Wrath* by J. Steinbek; *Martin Eden* by J. London; *The Great Gatsby* by F.S. Fitzgerald. The fragments with NC were retrieved from: *Martin Eden by* J. London; *A Farewell to Arms* by E.A. Hemingway; *Insurgent Mexico* by J. Reed; *The Invisible Man* by H.G. Wells. Studying phonosemantic relationships between prose texts with PC and NC and English initial consonant clusters resulted in the existence of significant correlation between the clusters *Fl-* ($\chi^2 = 20{,}82$), *Gl-* ($\chi^2 = 7{,}44$), *Pl-* ($\chi^2 = 7{,}18$) and *Sp-* ($\chi^2 = 4{,}72$) and texts with PC, on the one hand, and the clusters *Br-* ($\chi^2 = 3{,}96$), *Fr-* ($\chi^2 = 4{,}95$), *Str-* ($\chi^2 = 5{,}63$**)**, *Thr-* ($\chi^2 = 8{,}56$**)** and *Tr-* ($\chi^2 = 4{,}49$**)** and texts with NC, on the other hand.

After the application of the centering procedure, the average ratings of these clusters on the "evaluation" category appeared to be the following:

Fl- 0,6	*Br-* -0,3
Gl- 0,1	*Fr-* -0,6
Pl- 0,5	*Str-* -0,1
Sp- 0,1	*Thr-* -0,1
	Tr- -0,1

Similarly to the procedure with the poetic texts, we applied formula (2) to determine PhS of prose fragments with PC and NC. PhS of "pleasant" prose texts eguals:

$$PhS_{\chi 2} = (20{,}82 \times 0{,}6) + (7{,}44 \times 0{,}1) + (7{,}18 \times 0{,}5) + (4{,}72 \times 0{,}1) = 17{,}3$$

PhS for "unpleasant" prose fragments is:

$$PhS_{\chi 2} = (3{,}96 \times -0{,}3) + (4{,}95 \times -0{,}6) + (5{,}63 \times -0{,}1) + (8{,}56 \times -0{,}1) + (4{,}49 \times -0{,}1) = -6{,}03$$

Thus, PhS of prose texts with PC is larger than in the texts with NC. It has already been proven that the symbolic activity of texts with PC is larger than in the texts with NC. Consequently, the correlation between the meanings of certain clusters and the content of texts in prose is, at least according to one characteristic ("evaluation"), unquestionable. One can observe significant correlation between the "pleasant" clusters and the imaginative texts with PC; the texts with NC have a high frequency of "unpleasant" initial consonant clusters.

Determining phonetic symbolism of newspaper texts. Since newspaper texts are devoid of any emotionality and expressiveness – they mostly inform the reader but don't affect him – it is impossible to objectively "evaluate" them as "pleasant" or "unpleasant". But having counted the frequency of English initial consonant clusters in newspaper articles (*The New York Times*, heading *"National Politics"*) we found out that the clusters *Pr-*, *Sp-*, *Tr-*, *St-* and *Kw-* are the most frequent in newspaper style. Average ratings of these clusters according to the six categories of Osgood's differential were calculated by taking into account the centering method. The results are presented in Table 2.

Table 2
Average Ratings of the Most Frequent Consonant Clusters
according to the Semantic Differential

	Power	Evaluation	Activity	Evenness	Rigidness	Size
Pr-	0,5	0,2	0,1	0,1	0,4	0,1
Kw-	-0,2	-0,6	-0,4	-0,6	-0,2	-0,2
Sp-	0,4	0,1	0,3	0	0,03	0,2
St-	0,7	0,2	0,7	0,1	-0,2	0,2
Tr-	0,7	-0,1	0,2	-0,1	-0,2	0,1

It is evident that the category "power" is the only one out of the six where all the clusters possess the highest average ratings. We admit that this category is the most active and especially expressive in newspaper texts. From the results of our abovementioned psycholinguistic analysis of the sound-symbolic properties of 30 English initial consonant clusters, four out of five clusters are "strong". Only the cluster *Kw-* turned out to be weak, but even though this value is the largest among the values on the other categories, its phonosemantic correlation with the newspapers is the weakest out of five significant ones as *Kw-* $\chi^2 = 3{,}86$. Knowing the values of χ^2 in the newspaper genre for all five clusters (*Pr-* $\chi^2 = 425{,}01$; *Sp-* $\chi^2 = 18{,}5$; *Tr-* $\chi^2 = 14{,}9$; *St-* $\chi^2 = 5{,}9$; *Kw-* $\chi^2 = 3{,}86$) we may apply the same procedure used for poetry and prose (formula 2). This experiment aims to confirm the assumption about a "strong" or "weak" character of newspaper texts. If we multiply the χ^2 values of each cluster by its corresponding average rating on the category "power" and find the sum total of these products, the result will be as follows:

$$PhS_{\chi 2} = 212{,}5 + -0{,}77 + 7{,}4 + 4{,}13 + 10{,}43 = +233{,}69$$

The positive result shows that newspaper texts can be evaluated as "strong"; hence, as it was mentioned before, the results with the symbol "+" refer to positive poles whereas results with the symbol "-" refer to negative ones. But it is quite possible that newspaper texts are characterized by some other phonetic symbolic meanings which are inherent to the clusters in Table 2. That is why it is reasonable to make calculations for these texts according to all the categories of this table. Currently, there are no objective criteria for the distribution of publicist or newspaper texts according to their connotation; hence there is no explicit data as to the type of phonetic symbolism on semantic differential categories that can be characteristic of newspaper texts. We suggest applying formula (2) for determining PhS of newspaper texts according to the following categories:

	$PhS_{\chi 2}$
"evaluation"	+84,2;
"activity"	+53,62;
"evenness"	+39,28;
"rigidness"	+173,9;
"size"	+48,098;

Preliminary calculations showed that the largest result of *PhS* of newspaper texts was displayed in the category "power" (see above $PhS_{\chi 2} = +233{,}69$). Two more large results of *PhS* are "rigidness" ($PhS_{\chi 2} = +173{,}9$) and "evaluation" ($PhS_{\chi 2} = +84{,}2$).

Thus, we divided all the poetic and prose texts into two opposite groups regarding their connotative meaning, where the poles of the meaning "positive"

or "negative" were determined by the researcher. If the occurrence of phoneme combinations of the type *Bl-*, *Br-*, *Gr-* etc. in two groups of texts is of an arbitrary nature and the researcher was much mistaken in subdividing these texts into two opposite groups, then the distribution of frequencies of the analyzed clusters will be approximately proportional in different texts and will correspond to the theoretically expected values. However, the statistical analysis corroborated that the frequencies of certain clusters significantly exceed the expected theoretical values in one group of texts and the frequencies of the other clusters dominate in the other group of texts (χ^2 values indicate it). By itself, the high frequency of a sound or a sound combination in the texts doesn't necessarily prove the existence of symbolic meaning expressed by these sounds. Symbolic function of a sign emerges, as is shown in some scientific papers (Sieberer 1947; Levickij 1969), only when the sound form of a sign (or the sound level of the text) corresponds to its content.

Therefore the conclusions about the symbolic function of sounds based only on their high frequency in the text should be considered fallacious. The PhS of the sound (or sound combination) must *correspond* to the connotative or denotative meaning of the word or text. This very correspondence was determined in the result of our analysis. Statistically significant clusters with "pleasant" PhS occur in the texts with PC, and the clusters with "unpleasant" PhS appear in the texts with NC. All these facts no doubt prove existence of certain PhS in the analyzed texts, and the PhS is expressed through the initial consonant clusters. But this conclusion can not be made concerning the newspaper texts as it is impossible to subdivide them into any groups based on the connotative meaning (at least those texts selected for our analysis). Thus, the PhS values obtained for these texts may be interpreted as possible but not finally determined. Additional experiments based on different kinds of publicist texts that are written in different languages need to be carried out. Only then can one assume that publicist texts also possess PhS. Even if this assumption is corroborated one should remember the nature of PhS in such texts: Phonetic symbolism of publicist texts is more hidden and latent than that of the poetic texts, for instance, where it is not always explicitly expressed.

References

Abelin, A. (1999). *Studies in sound symbolism*. Gőteborg: Department of Linguistics, Gőteborg University.

Brown, R., **Nuttal, R.** (1959). Methods in phonetic symbolism experiments. *Journal of Abnormal and Social Psychology 59, 388-389.*

Hinton, L., Nichols, J., Ohala, J. (1994). *Sound symbolism.* Cambridge University Press.

Kronasser, H. (1952). *Handbuch der Semasiologie*. Heidelberg: J. Groos Verlag.

Kushneryk, V.I., Levickij, V.V. (1986). Phoneticheskaya motivirovannost i foneticheskoe znachenie slova. In: *Psihologicheskie problemy semantiki i ponimaniya teksta: 21-27.* Kalinin.

Levickij, V.V. (1998). *Zvukovoj simvolizm. Osnovnye itogi*. Černovcy: Ruta.

Levickij, V.V. (1969). K probleme zvukovogo simvolizma. In: *Psihologicheskie i psiholingvisticheskie problemy vladeniya i ovladeniya yazykom: 123-132.* Moskva: MGU.

Lowrey, T., Shrum, L. (2007) *Phonetic symbolism and brand name preference.* http://business.utsa.edu/marketing/files/phdpapers/Tina1-JCR2007.Final.pdf

Lvova, N.L. (2004). Pro symvoliku pochatkovyh prygolosnyh spoluchen v suchasnij anglijskij movi. *Naukovyj visnyk 188/89, 47-61.* Černovcy: Ruta.

Magnus, M.A. (1999). *Dictionary of English sound*. http://www.trismegistos.com/

Shtern, A.S. (1969). Objektivnye kriterii vyyavleniya effekta "zvukovoi simvoliki". In: *Materialy seminara po probleme motivirovannosti yazykovogo znaka: 69-73.* Leningrad: LGU.

Sieberer, A. (1947). Primäre oder sekundäre Lautbedeutsamkeit? *Anzeiger der Österreichischen Akademie der Wissenschaften, Philosophisch-historische Klasse 9, 35-52.*

Taylor, I.K., Taylor, M.M. (1965). Another look at phonetic symbolism. *Psychological Bulletin 64(6), 413-427.*

Tsuru, S., Fries, H.S. (1933). A problem in meaning. *Journal of General Psychology 8, 28-284.*

Ullmann, S. (1964). *Semantics: An introduction to the science of meaning*. Oxford: Basil Brackwell.

Zhuravliov, A.P. (1974). *Phoneticheskoe znachenie*. Leningrad: LGU.

Kelih, E. et al. (eds.), Issues in Quantitative Linguistics, Vol. 2, 104–109

A combinatorial method for context comparison

Jiří Milička

Introduction

Different branches of linguistics agree on the idea that the context of a word is a clue to its semantics and the way of its usage. Especially corpus linguistics insists on the key role of context in language research.

This paper suggests a model on which we can base a method for testing various hypotheses related to context. For example we assume that L1 context of some word type "is similar" to L1 context of another one. We can easily determine these two contexts: We just find all occurrences of the chosen types and gather their L1 contexts in two groups. But how can these two groups be compared? What algorithm should we choose? And how "dissimilar" have these groups to be to falsify our assumption?

Within applied linguistics, this task is quite common and we can describe it in more general terms: we choose a group of tokens with the feature we wish to explore; from the rest of the tokens left in the text, we choose another group of tokens according to a different feature. Now we compare these two features by comparing the two groups of tokens. The feature may be the context, style, authorship, conditions of production etc. Many statistical metrics were created for a comparison like this (often on an ad hoc basis, e.g. see Cvrček 2010) and therefore the results of such researches are hardly comparable and sometimes difficult to interpret.

We suggest a comparison method that is universal and easy to interpret, but due to its computational complexity[1] it will probably not play an important role within NLP applications, finding its use rather in testing quantitative linguistic theories.

Our methodology and epistemology pick up the threads of our article published in *Glottotheory* (Milička 2009). While a typical quantitative linguistic research would first quantify and measure some feature of a text and then try to find a regression or a general hypothesis fitting the data, we approach the real data less straightforwardly. In the article mentioned above, we asked ourselves whether the type-token relation (TTR) curve tells us anything about the structure

[1] However, an algorithm based on this method is easy to parallelize and so the contemporary CPU development which tends to integrate more cores into one chip rather than increase its speed does not present a problem.

of the text given. The question could not be answered before finding the nature of TTR when measured on a "text without a structure". Or rather the curve of an average TTR measured for all permutations of the text. The difference between this curve and TTR curve of a real text is related to the structure of the text. Of course we did not actually permute the text, but we algebraically derived a model that describes the result of such permutations. Now we ask ourselves analogically, how "similar" would an average result be if we chose the groups from all permutations of the text.

Derivation of the model

Let us imagine a text.[2] For example this children's rhyme:

> One little two little three little Indians
> Four little five little six little Indians
> Seven little eight little nine little Indians
> Ten little Indian boys.

As the first group of the tokens (*A*) we choose the first line of the song, as the second group (*B*) we choose the third and the fourth lines.[3]

(*A*)
One **little** two **little** three **little Indians**

(*B*)
Seven **little** eight **little** nine **little Indians**
Ten little Indian boys

We see that four tokens are matching[4]. And we ask whether the number is large or small. How many tokens would be matching on average if we examined all possible pairs of samples (with the same length as the subsets *A* and *B*, that is 7 and 11 tokens) from the given text? Or, to put it differently, how many tokens would be matching on average if we chose such pairs from all permutations of the text?

---

[2] By terms of the multiset theory: Let us imagine a text as a multiset *T*, where set of all types in the text is the *underlying set of elements* and the absolute frequency of each type is its *multiplicity*

[3] We choose the submultisets *A* and *B* so that $A \uplus B \sqsubseteq T$.

[4] The number of the matching tokens (*g*) is equal to the cardinality of the conjunction of these two submultisets $g = A \cap B$.

Let us start with the question, what is the probability r_i that type I (that occurs f_i times in the text T) would occur exactly m times in subset A and n times in subset B. For each token of the text we can determine, whether it belongs to type I or not – this is a binary decision. Hence the number (p_0) of all the possibilities of how the tokens of the type I can be transposed in the text is equal to the number of multiset permutations of two elements, the multiplicity of the first is equal to f and the multiplicity of the second equals $|T| - f$.

$$p_0 = \binom{|T|}{f_i}$$

Now we determine the number ($p_{m,n}$) of permutations that fit the condition that m tokens of type I are in group A, n tokens in group B and the rest of the tokens of type I are in the rest of the text.[5] We then multiply the following numbers: the number of all the possibilities of how m tokens can be transposed in $|A|$ positions, n tokens in $|B|$ positions and the rest of the tokens of the type I ($f_i - m - n$) in the rest of the positions ($|T| - |A| - |B|$).

$$p_{m,n} = \binom{|A|}{m}\binom{|B|}{n}\binom{|T| - |A| - |B|}{f_i - m - n}$$

The wanted probability r equals the ratio of the number of permutations that fit the condition mentioned above to all the permutations of the type.

$$r_i = \frac{\binom{|A|}{m}\binom{|B|}{n}\binom{|T| - |A| - |B|}{f_i - m - n}}{\binom{|T|}{f_i}}$$

And now we sum up the ratio (r_i) for all the types occurring in the text and for all possible m and n.

$$g = \sum_{i=2}^{|Supp\ T|} \sum_{m=1}^{|A|} \sum_{n=1}^{|B|} \min(m,n) \frac{\binom{|A|}{m}\binom{|B|}{n}\binom{|T| - |A| - |B|}{f_i - m - n}}{\binom{|T|}{f_i}}$$

$m + n \leq f_i \wedge f_i - m - n \leq |T| - |A| - |B|$

If $A \uplus B = T$ then the formula could be simplified:

[5] I.e. the multiplicity of the element I in the submultiset A is equal to m and the multiplicity of the element I in the submultiset B is equal to n and the multiplicity of the element I in $T \ominus A \ominus B$ is equal to $f_i - m - n$.

$$g = \sum_{i=2}^{|Supp\ T|} \sum_{m=1}^{\min(|A|, f_i - 1)} \min(m, f_i - m) \frac{\binom{|A|}{m}\binom{|T| - |A|}{f_i - m}}{\binom{|T|}{f_i}}$$

The following table shows how the model fits the real data. The figures in the third column present the average cardinality of the intersection of *A* and *B*, randomly selected million times from The Last of the Mohicans by James Fenimore Cooper. The figures in the fourth column were calculated according to the formula.[6]

|A|	**|B|**	**Real text**	**Model**
1000	1000	505,598	505,586
150	100	34,6007	34,6000
50	25	6,05841	6,05592
20	5	0,81816	0,81827
2	2	0,05180	0,05202
10 000	10 000	7333,08	7333,07
1000	3	1,82223	1,82069

An application example

Let us demonstrate the use of the model for testing a hypothesis:

> The immediate right context (R1) of the word *say* is more similar to the immediate right context of the word *says* than the right context of the word *said*.

For testing the hypothesis we use The Last of the Mohicans by James Fenimore Cooper:

[6] The program we used and its documentation are available at www.milicka.cz/context.

Type	Frequency
say	567
says	64
said	991

Pair	Model	Real text	Ratio
say – says	35,02	47	1,34
say – said	324,44	292	0,90

This hypothesis could be generalized and tested for all verbs in a larger corpus: the R1 context of a verb in the form without any ending is on average more similar to the R1 context of a 3rd person sg present form than to a past form R1 context.

A metrics that would determine the probability that e.g. the R1 context of the pair *say – said* would have 292 or less matching tokens would be of great use for testing such hypotheses as it would enable us to develop statistical tools similar to Student's Test. We leave this for further consideration.

Conclusion

As mentioned above, we can use the proposed model for various purposes like comparing sets of words chosen according to different mutually exclusive criteria (i.e., we cannot compare the L1 context of a word with the P1 context of another word, because we could happen to include one token into both the sets compared.

Quantitative linguistics is not the only field for the model to find its use. We can imagine it being used within other branches of science.[7]

[7] E.g. to test a biological hypothesis that in beech crowns there nest similar bird species as in the crowns of oaks.

References

Cvrček, V. (2010). A contextual approach to parts of speech. In: *Intercorp: Exploring a Multilingual Corpus*: *190 – 204*. Prague: NLN.

Milička, J. (2009). Type-Token & Hapax-Token relation: A combinatorial model. *Glottotheory 2/1*, 99 – *111*.

Syropoulos, A. (2001). Mathematics of multisets. In: *Multiset Processing*: *347–358*. London: Springer-Verlag.

Kelih, E. et al. (eds.), Issues in Quantitative Linguistics, Vol. 2, 110-116

Thematic concentration in texts

Ioan-Iovitz Popescu
Gabriel Altmann

Introduction

As is well known, data are not given but constructed. Their proto-image is a concept which is in turn projected in some (at least intuitive) operational form on the reality in which material specimens of the concept are found. These specimens represent data. Data are constructed for a special aim; hence they are neither true nor false but rather adequate or not adequate. They are the basis of classification, orientation in reality, and in sciences, they are created in order to test hypotheses. The history of science is a history of changing and improving concepts.

Here we shall be concerned with a very abstract concept of thematic concentration of a text which has been introduced in Popescu et. al. (2009, Chapter 6) as a normalized sum of weighted ranked frequencies of those autosemantic words whose rank is smaller than or equal to the *h*-point. The procedure of computation is as follows: First the frequencies of word forms in a text will be counted and ranked according to decreasing magnitude. Then for the given rank-frequency distribution the *h*-point will be computed according to formula:

$$h = \begin{cases} r_i & \text{if there is an } r_i = f(r_i) \\ \dfrac{f(r_1)r_2 - f(r_2)r_1}{r_2 - r_1 + f(r_1) - f(r_2)}, & \text{if there is no } r_i = f(r_i) \end{cases} \tag{1}$$

If there is no r_i which is equal to $f(r_i)$ one takes the two respective neighboring values of r such that $r_1 < f(r_1)$ and $r_2 > f(r_2)$ – usually $r_2 = r_1 + 1$ – and computes h according to the second row in (1). In the next step one takes into account only those autosemantics whose rank is smaller than or equal to the integral part of h and their ranks are marked as r'. The formula for thematic concentration takes into account both the distance of r' from h and its weight $f(r')$, sums the results and normalizes them by the maximum that could be attained. Thus one obtains

$$TC = 2\sum_{r'=1}^{T} \frac{(h - r')\, f(r')}{h(h-1)\, f(1)}, \tag{2}$$

where T is the number of autosemantics whose rank is smaller than h. Another form of evaluation has been proposed in Popescu, Altmann (2008).

In spite of this simple operationalization the question remains whether word forms yield an appropriate set of data suitable for expressing text concentration. It is not only the concept of a phenomenon that may vary, the data may be different, too, and our first duty is to look for data which adequately capture our conceptual constructions.

Of course, the number of possibilities to quantify thematic concentration is not limited. For example the study of sentences associated by a reference is frequently used but it produces clumsy graphs whose properties must be evaluated; it is rather non-transparent for long texts. Up to now the above mentioned method is the only one based on simple principles. Our task is only the finding of appropriate data. In the next section we shall describe some possible approaches.

Three approaches

The first approach, as described above, takes into account all word forms. Since in strongly synthetic languages many words have different forms, some autosemantics repeated in the text several times do not obtain a rank smaller than *h* even if they are relevant for the theme. Consequently, texts in synthetic languages would display smaller thematic concentration. This fact could, of course, be used for typological purposes; but here we are interested in individual texts, not in languages.

The second approach leads (quasi) logically to the lemmatization of the text. Thereby all forms of a word are unified and the frequencies of forms are added. Thus thematically relevant autosemantics may obtain ranks smaller than *h*. However, in strongly synthetic languages, forms of auxiliaries may be unified and thereby obtain smaller ranks, too. This case can be demonstrated using the poem *Der Erlkönig* by J.W.v. Goethe. The word forms and their frequencies up to rank 10 are presented in Table 1.

Table 1
Frequencies of first ten word forms in Goethe's Erlkönig

Rank	Word Form	Frequency
1	mein	11
2	und	9
3	Vater	9
4	du	7
5	mit	6
6	es	5
7	Kind	5
8	er	5
9	den	5
10	so	4

The *h*-point is here $h = 5.5$ following from (1) as $[6(6)-5(5)]/2 = 5.5$ and there is only one autosemantic word (*Vater*) whose rank is smaller than 5.5. Thus we obtain

$$TC = 2(5.5 - 3)9/[5.5(4.5)11] = 0.1653$$

Since *Vater* has the same rank as *und*, the exchange of ranks leads to

$$TC = 2(5.5 - 2)9/[5.5(4.5)11] = .0.2314.$$

Taking the mean we obtain $TC = 0.1984$ which is equal with TC if we ascribe both *Vater* and *und* the same mean rank 2.5

However, if we lemmatize the poem, we obtain the results presented in Table 2. As can be seen, the pre-*h* domain now contains more auxiliaries and synsemantics, but there is again only one autosemantic word (*Vater*) Here $h = 6.75$ and

$$TC = 2(6.75 - 6)9/[6.75(5.75)24] = 0.0145$$

If we place *Vater* at rank 5 (because of equal frequency with *und*), we obtain $TC = 0.0338$. In order to obtain a unique measure we can take the mean of these two values, i.e. $(0.0145 + 0.0338)/2 = 0.0242$ which is the same as that obtained with averaged ranks. The difference between 0.1984 and 0.0242 is too great. Thus lemmatization reorganizes the whole frequency field by eliminating the synthetism. Hence using it for expressing thematic concentration measured in this way must be treated with caution. If we compare different languages, we subtract different degrees of synthetism from both. Nevertheless, it can be used for measurement of synthetism and for typology. It represents a mid position between the word-form approach and the referential approach which will be shown in the sequel.

Table 2
Frequencies of the first ten lemmas in Goethe's Erlkönig

Rank	Lemma	Frequency
1	ich	24
2	er	14
3	du	13
4	der/die/das	13
5	und	9
6	Vater	9
7	sein (verb)	6
8	mit	6
9	Erlkönig	5
10	Kind	5

We know that many synsemantics are merely references to the main words or phrases having a main word, and verbal endings may refer to persons, too; persons are called differently like in *Erlkönig: Knabe, Sohn, Kind* represent the same person, some cases of *du, er, es* refer to the same person, etc., hence a third way would be the finding of *hrebs* (cf. Hřebíček 1992, 1993, 1995, 1997), i.e. sets of entities referring to the same textual or real objects. In that case pronouns are always elements of the set representing a noun or a nominal phrase. In Table 3 we present the hreb-analysis of Erlkönig on the word level – though there are other possibilities, too (cf. Köhler, Naumann 2007). The table is taken from Ziegler, Altmann 2002: 32). For the sake of simplicity we enumerate all occurrences of the elements of a hreb in the same order as they occur in the poem. As can be seen, some pronouns occur in different hrebs because in the given case they have different referents.

Table 3
Hreb analysis of Goethe's Erlkönig, the first seven ranks

Rank	**Hreb**	**Frequency**
1	Kind = [Kind, Knaben, ihn, ihn, Sohn, du, dein Sohn, du, Kind, dir, mein, mein, mir, Kind, Knabe, du, dich, dich, mein, mein, Sohn, Sohn, dich, deine, du, mein, mein, mich, mir, Kind, Kind]	32
2	Vater = [wer, Vater, seinem, er, er, er, mein, Vater, du, mein, Vater, Vater, du, mein, Vater, Vater, du, mein, mein, ich, Vater, Vater, Vater, er, er, seinen]	26
3	Erlkönig = [Erlkönig, Erlenkönig, mir, mich, meine, Erlenkönig, mir, mein, meine, Erlkönigs, ich, mich, ich, er, Erlkönig	15
4	der = [der, den, dem, den, den, dem, der, den, dem, den]	10
5	und = [und, und, und, und, und, und, und, und, und]	9
6	ist = [ist, ist, sind, sei, bist, war]	6
7	mit = [mit, mit, mit, mit, mit, mit]	6

Here we have exactly $h = 6$ but in the pre-h domain the first three hrebs contain thematic autosemantics, hence we obtain

$$TC = \frac{2}{6(5)32}[(6-1)32 + (6-2)26 + (6-3)15] = 0.6438,$$

a value that better corresponds to our intuitive image of thematic concentration in the given poem. The poem has three main objects and the hreb-analysis makes them visible.

In two other texts processed by Ziegler and Altmann (2002) by hreb-analysis, namely Goethe's poem *Epiphanias* and in E. Strittmatter's prosaic work *Der Erdstern*, we find *TC* = 0.77 and *TC* = 0.48 respectively. A very preliminary judgement may be the statement that poetry is more concentrated than prose. However, this is merely a starting hypothesis for a thorough investigation of different text sorts.

The asymptotic variance of *TC* can be computed as follows

$$(3) \qquad Var(TC) = \left[\frac{2}{h(h-1)f(1)}\right]^2 Var\left(\sum_{r'=1}^{T}(h-r')f(r')\right).$$

Let $\Sigma f(r') = n$, then

$$Var(TC) = C\left[Var\left(hn - \sum_{r'=1}^{T} r' f(r')\right)\right].$$

Since *hn* is a constant and $\Sigma r' f(r') = nm_{1,r'}$ we obtain $Var(TC) = C[Var(nm_{1,r'})] = Cn^2 m_{2,r'}/n$ and finally

$$(4) \qquad Var(TC) = \left[\frac{2}{h(h-1)f(1)}\right]^2 nm_{2,r'}$$

where $m_{2,r'}$ is the second central moment of the *T* autosemantic ranks. To note, $m_{1,r'}$ is the mean of *r'* ranks. For example, from Table 3 we obtain

$$m_{1,r'} = [1(32) + 2(26) + 3(15)]/73 = 1.7671$$

hence

$m_{2,r'} = [(1 - 1.7671)^2 32 + (2 - 1.7671)^2 26 + (3 - 1.7671)^2 15]/73 = 43.0411/73 = 0.5896.$

Thus the variance of *TC* for Table 3 is

$$\text{Var}(TC\text{-Table 3}) = 4(73)(0.5896)/[6^2(5)^2 32^2] = 0.0001868$$

Using this computation an asymptotic test for the difference of two texts can be constructed. Comparing the hreb-like analyses of *Erlkönig* and Strittmatter's *Der Erdstern* for which we obtain *Var*(*TC-Erdstern*) = 0.00000107, the test yields

$$u = (0.64 - 0.48)/\sqrt{(0.0001868 + 0.00000107)} = 11.67$$

signalizing a highly significant difference.

Though the hreb-like analysis seems to optimally express the thematic concentration, it can be presented in different variants. For example, if we consider the article as an integer part of the noun – some languages do not have articles at all, other ones join them with the noun – or if we consider only phrases, the rank frequency distribution would change and consequently TC, too. Word-like hrebs represent an acceptable analysis but one could also go a step deeper and take into account also the person and the object of conjugation which create further references, for example in Hungarian the verb forms *látok, látom, látlak* all mean "I see", i.e. *látok* expresses reference to the first person only but it does not contain any other reference, the second, *látom*, refers to a definite object or to the reverential form of the second person (*magát, önt*) and the third refers to the second person. Thus, performing the hreb-like analysis the researcher must state which basic definitions he used, otherwise comparisons are not possible. One can strictly reject agreement or accept it, etc.

Problems to be solved in the future are numerous. First, which kind of counting yields the "best" results? But in order to define "best", one must have an a priori hypothesis. Thus setting up hypotheses would be the next step. This could be accomplished by examining the relationship of thematic concentration to other text properties. Presently, no decision can be made but preliminarily we can accept the criterion that that measurement of thematic concentration which displays the maximal number of interrelations with other properties is the most adequate one.

References

Hřebíček, L. (1992). *Text in communication: supra-sentence structures*. Bochum: Brockmeyer.

Hřebíček, L.(1993). Text as a construct of aggregations. In: Köhler, R., Rieger, B. (eds.), *Contributions to quantitative linguistics*. Dordrecht: Kluwer.

Hřebíček, L.(1995). *Text levels. Language constructs, constituents and the Menzerath-Altmann law*. Trier: WVT.

Hřebíček, L.(1997). *Lectures on text theory*. Prague: Oriental Institute.

Köhler, R., Naumann, S. (2007). Quantitative analysis of co-reference structures in texts. In: P. Grzybek und R. Köhler (eds.), *Exact Methods in the Study of Language and Text*: *317-330.* Berlin-New York: de Gruyter.

Popescu, I.-I., Altmann, G. (2008). Autosemantic compactness of texts. In: Altmann, G., Zadorozhna, I., Matskulyak, Y. (eds.), *Problems of General, Germanic and Slavic Linguistics. Papers for 70-th anniversary of Professor V. Levickij: 472-480.* Chernivtsi: Books – XXI.

Popescu. I.-I. et al. (2009). *Word frequency studies*. Berlin-New York: Mouton de Gruyter.

Ziegler, A. (2005). *Denotative Textanalyse*. In: Köhler, R., Altmann, G., Piotrowski, R. G. (Eds.), *Quantitative Linguistics. An International Handbook: 423–446*. Berlin-New York: de Gruyter.

Ziegler, A., Altmann, G. (2002). *Denotative Textanalyse*. Wien: Praesens.

Kelih, E. et al. (eds.), Issues in Quantitative Linguistics, Vol. 2, 117-128

Complexity of the Vai script revisited: A frequency study of the syllabary

Andrij Rovenchak
Charles Riley
Tombekai Sherman

0. Introduction

The present paper is a continuation of the quantitative studies of writing systems, in particular in the domain of African indigenous scripts.

We analyze the statistical behavior of the script complexity defined according to the composition method suggested by Altmann (2004). The analysis of the Vai script complexity was made in a recent paper (Rovenchak, Mačutek, Riley 2009). To recall briefly, the idea of this approach is to decompose a letter into some simplest components (a point is given the weight 1, a straight line is given the weight 2, and an arc not exceeding 180 degrees has the weight of 3 units). The connections of such components are: a crossing (like in X) with weight 3, a crisp (like in T, V or L) with weight 2, and the continuous connection (like in O or S) with weight 1. For the Vai script, we also suggested that filled areas are given the weight of 2. With this method, a number of scripts had been analyzed: Latin (Altmann 2004), Cyrillic (Buk, Mačutek, Rovenchak 2008), several types of runes (Mačutek 2008), Nko (Rovenchak, Vydrin 2010).

The uniformity hypothesis for the distribution of complexity was confirmed for all the scripts but the Vai syllabary. The failure in the latter case can be caused by the fact that syllabaries require some modification of this hypothesis as all the other scripts analyzed so far are alphabets.

Here, it is worth mentioning the definitions of the discussed script types (Daniels 1990). In alphabets, a character (letter, symbol) denotes mostly one sound, either a consonant or a vowel. Some scripts do not mark all the vowels (e.g., in Arabic, only long vowels are written), and this type of scripts can be referred to as *abjad*. In a syllabary, a character denotes mostly a combination of the "consonant + vowel" type. The best known examples are Japanese *kana*, namely *hiragana* and *katakana* or the Cypriot syllabary. Slightly different approach is used in the Indic scripts derived from Brahmi, where special modifiers are added to change an inherent vowel associated with a particular symbol, hence such scripts are referred to as alphasyllabaries. Such an approach is also known for the Amharic script *fidel* having a complicated genealogy. This very script gave the name for this script type, *abugida*, from the standard letter order. Note that, for instance, Japanese *kana*, the Cypriot script, and native Cherokee script are pure

syllabaries as the shapes of symbols with different vowels are unrelated. The same applies to the Vai script, an indigenous writing system of the Vai people (Liberia) originated in the 1820s.

After its invention, the Vai script continued in use through the remainder of the 19th century and into the 20th. Archives of manuscripts were kept at the villages of Jondu and Bandakolo, but both were burned in raids by the neighboring Golas around the turn of the century (Dalby 1967). Four manuscripts surviving from the nineteenth century are known to exist or were copied. In 1913, a 180-page manuscript by Boima Kiakpombo was produced, kept in diary form. Momolu Massaquoi, consul to Hamburg, produced translations in Vai script of some religious texts and began to collaborate with August Klingenheben, a German linguist. Klingenheben's involvement with Vai culminated in a collaboration on a standardized form of the script, worked out at the University of Liberia with elders from several towns of the Vai country in 1962. Jangaba Johnson, Bai T. Moore, and Mohamed Nyei also were active throughout the latter half of the 20th century in working with Vai. In 2003, with the assistance of SIL and Lutheran Bible Translators, a New Testament in Vai was produced, with Tombekai Sherman serving as chair of the translation committee. Sherman continued to produce texts in Vai and, with Mohamed Nyei, S. Jabaru Carlon and others, contributed toward the standardization of the Vai script into Unicode.

1. Vai texts

We have analyzed four Vai texts. They are:

1) *The Universal Declaration of Human Rights* (UDHR);
2) *Vai Proverbs and Rhymes* (VPR);
3) *Our Village* (OV);
4) *Who Were the Vai?* (WWV).

The Universal Declaration of Human Rights (UDHR) was written under the authority of a standing body of the United Nations, the Commission on Human Rights, consisting of Eleanor Roosevelt (chair), with contributions from John Peters Humphrey, René Cassin, Charles Habib Malik, Chang Peng-chun, and other members of the Commission. On December 10, 1948, it was adopted by the UN General Assembly with no dissent, and only eight abstentions. Its translation into Vai was prepared by Tombekai Sherman.

Vai Proverbs and Rhymes were compiled in 2008 by Tombekai Sherman, the list contains over 200 proverbs supplied with some introductory information and comments.

Our Village is a narrative written by Tombekai Sherman.

Who Were the Vai? is an article published in the *Journal of African History* by Adam Jones (1981). Its translation into the Vai language, using the Vai script, was undertaken by Sherman between 2007 and 2009 with the permission of Cambridge University Press.

The abovementioned texts are homogeneous with respect to the orthography which in an important issue for the study of texts written in indigenous scripts.

2. Preliminary notes

Our previous study of the complexity of the Vai script revealed that the distribution of complexity values does not conform to the uniformity hypothesis. Comparing to alphabets, the set of characters in syllabaries is typically several times larger. The size of this set significantly depends on the phonotactics of a particular language and on the approach used to map the phonetic structure onto the written representation. For instance, the Cherokee syllabary has 85 characters, the Vai syllabary has over 200 characters, but the modern Yi script contains about 800 signs as the tonal distinction are implemented as separate shapes.

We suggest dividing the symbols in the syllabary into two parts, *core* (containing the most frequent characters) and *periphery* (correspondingly, all the remaining characters). There is no unique way to define the number of characters in the core, and we propose to use the notion of *h*-point (Popescu et al. 2009) in order to separate the syllabary.

The h-point definition. Given a set of data sorted in a descending order with respect to absolute frequencies, assign the rank $r = 1$ to the most frequent item, then rank $r = 2$ to the next most frequent, etc. Let the frequency of the r-ranked item be $f(r)$. The h-point is defined as the solution of the equation

$$r_h = f(r_h), \tag{1}$$

so that all the items with frequencies higher than $f(r_h)$ have ranks lower than r_h and vice versa.

This definition is directly related to the so-called h-index suggested by Hirsch (2005) to measure the output of a scientist. A similar quantity was proposed several decades earlier by Sir Arthur Eddington for the estimation of the achievements of long-distance cyclists (Barrow 2002, p. 83; E Numbers 2008).

It is possible that in some sample no frequency satisfies Eq. (1). In such a case, the value of the h-point can be defined by simple interpolation between the neighboring values corresponding to ranks r_1 and r_2 as follows:

$$(2) \quad r_h = \left\lceil \frac{r_2 f(r_1) - r_1 f(r_2)}{r_2 - r_1 + f(r_1) - f(r_2)} \right\rceil,$$

As both rank and absolute frequency are integers, we apply the ceiling function $\lceil x \rceil$ equal to the next integer after x.

When applied to the word frequencies, the *h*-point is believed to divide the word list into mostly synsemantic (auxiliary) and autosemantic (full-sense words) branches (Popescu et al. 2009). The relation between these parts of a dictionary seems however more complicated and probably cannot be defined solely by the *h*-point (cf. Buk 2010). Still, the *h*-point can be seen as separating a rank–frequency distribution into two regimes, and this very property we use in our work.

In the present work, we analyze the frequency distribution of syllabic characters of the Vai script and propose that the *core part* of the syllabary is composed of the characters having frequencies above and equal to the *h*-point. Note that such a definition means that different cores would be obtained from different samples, and only a large corpus of texts would allow stating some *gold-standard core*.

3. Frequency data and core part of the syllabary

We have compiled frequency lists for characters for all the Vai texts mentioned in the previous section. The frequency lists are given in Table 1.

Table 1
Frequency of the syllabic characters in Vai texts

r	OV				UDHR				VPR				WWV			
	Vai	f_r	C_r	PhT	Vai	f_r	C_r	PhT	Vai	f_r	C_r	PhT	Vai	f_r	C_r	PhT
1		365	16	a		361	16	a		330	24	i		190	13	lɔ
2		257	19	ŋ		265	20	nu		302	19	ŋ		164	8	la
3		252	20	nu		259	25	ha		233	16	a		161	17	ma
4		247	11	mu		218	11	mu		216	25	ha		156	19	ŋ
5		214	24	i		212	24	i		205	17	ma		116	24	i
6		206	17	ma		201	19	ŋ		195	26	e		111	23	ko
7		199	25	ha		190	17	ma		190	13	wa		74	32	ɛ

r	OV				UDHR				VPR				WWV			
	Vai	f_r	C_r	PhT	Vai	f_r	C_r	PhT	Vai	f_r	C_r	PhT	Vai	f_r	C_r	PhT
8		188	13	wa		166	12	mɔ		169	8	la		73	13	wa
9		162	13	lɔ		159	14	an		162	23	ya		71	18	fɛ
10		141	32	ɛ		145	13	lɔ		141	12	mɔ		65	20	nu
11		138	8	la		144	8	la		119	13	lɔ		62	10	na
12		135	18	kɛ		143	13	wa		117	20	le		59	25	ha
13		106	26	e		132	28	ɔ		109	11	mu		57	26	e
14		103	14	an		119	23	ko		104	9	ku		56	14	an
15		103	5	ɓɛ		117	23	ya		100	15	ti		52	23	ya
16		95	12	mɔ		110	26	hin		96	18	kɛ		51	16	a
17		93	23	ya		107	32	ɛ		96	11	ɓe		50	28	ɔ
18		89	23	ko		106	9	ku		87	10	na		50	7	ka
19		85	10	na		103	18	kɛ		86	14	an		48	18	kɛ
20		84	20	le		100	33	sa		86	23	ko		46	33	si
21		77	24	ɓa		96	15	ti		85	26	hin		40	9	ku
22		76	25	nɗa		92	5	ɓɛ		81	32	ɛ		40	26	kɔ
23		71	28	ɔ		79	10	na		81	5	ɓɛ		39	9	ki
24		67	24	nɗɔ		76	26	e		79	28	yɛ		30	20	lo
25		66	15	ti		67	26	ɲa		77	7	ka		29	28	va
26		65	20	ja		65	20	lo		70	20	nu		28	6	lɛ
27		64	7	ka		63	24	ɓa		66	22	wo		28	22	wo
28		63	33	sa		62	20	ja		64	28	ɔ		27	15	ti
29		62	26	hin		62	11	gbi		63	20	lo				
30		59	22	wo		61	14	tɔ		61	20	ja				
31		56	27	ɛn		60	14	ta		56	24	ɓa				

r	OV				UDHR				VPR				WWV			
	Vai	f_r	C_r	PhT	Vai	f_r	C_r	PhT	Vai	f_r	C_r	PhT	Vai	f_r	C_r	PhT
32		55	11	ɓe		59	22	sɔ		56	25	nɗa				
33		47	22	sɔ		59	7	ka		56	26	kɔ				
34		44	30	bi		54	28	yɛ		46	22	o				
35		43	8	mɛ		54	6	lɛ		46	26	ɲa				
36		42	9	ku		53	27	ɛn		45	6	to				
37		39	14	ta		52	20	le		44	33	sa				
38		38	11	gbi		48	18	fɛ		44	7	lu				
39						45	19	ɓo		44	22	sɔ				
40						37	18	nɗe		43	27	ɛn				
41										43	12	kpa				
42										42	13	ji				
r_h	38				40				42				28			
% till r_h	79.1				85.1				79.4				63.6			
Total	5432				5407				5587				3102			

Legend: r — rank in the frequency list; f_r — absolute frequency of the character; C_r — complexity of the character; PhT — phonetic value of the character (vowel nasalization is marked by /n/; according to the conventions typical for African linguistics, /y/ denotes IPA [j] and /j/ denotes IPA [ʤ]); % till r_h — the percentage of the characters from the core syllabary relative to all occurring characters (given in the Total row).

All the studied texts but WWV have almost equal number of syllabic characters. As expected (cf. Popescu et al. 2009:19), this reveals close estimations for the *h*-point of all three texts and a smaller value for the shorter WWV text. The same applies to the percentage of text covered by the characters from the core.

From the frequency point of view, the proportion of 80% roughly corresponds to an English text lacking ‘j’, ‘x’, ‘q’, and ‘z’ (cf. Lewand 2000 or any other statistics on English letter frequencies). For the illustration, in the remaining part of this paragraph the te#t is typed without these letters substituting them with a single number sign ‘#’. It is easily seen that such a modification does not cause any problems in understanding. An obvious reason is the low fre#uency of the mentioned letters (less than 0.2 per cent). Even #ointly, the four least-fre#uent

letters occur less than the next, ranked 22nd letter ‘k’. Note however that omitting a letter in an English te#t introduces probably less confusion than omitting a syllabic sign in a Vai one as words in Vai contain 2–3 syllables on average.

The cumulative list of characters from all the four texts consists of 55 symbols (occurring at least in one core part). These characters are listed in Table 2 ordered according to the Unicode values.

Table 2
Cumulative list of characters from the core parts of the syllabary from different texts

Vai	*C*	PhT	Vai	*C*	PhT	Vai	*C*	PhT	Vai	*C*	PhT	Vai	*C*	PhT
	26	e		9	ki		33	sa		20	lo		26	kɔ
	11	ɓe		16	a		20	ja		23	ko		12	mɔ
	20	le		14	an		23	ya		7	lu		32	ɛ
	18	nɗe		25	ha		7	ka		9	ku		27	ɛn
	24	i		13	wa		17	ma		11	mu		5	ɓɛ
	26	hin		24	ɓa		10	na		20	nu		18	fɛ
	30	bi		28	va		26	ɲa		28	ɔ		6	lɛ
	11	gbi		12	kpa		22	o		14	tɔ		28	yɛ
	15	ti		14	ta		22	wo		13	lɔ		18	kɛ
	13	ji		8	la		19	ɓo		24	nɗɔ		8	mɛ
	33	si		25	nɗa		6	to		22	sɔ		19	ŋ

From the point of view of a literate native speaker in the Vai script, one does not need to know a whole set of characters to be able to read and write using the syllabary. When literacy programs were intensified among Vai people, it took between two and three months for most people to be able to read most of the literature that was produced at that time since there were not so many characters involved.

One can compare this cumulative list to the repertoire of characters found in the *Book of Ndɔlɛ* (known also as the *Book of Rora*), the earliest known Vai manuscript of ca. 1850 (Stewart 1972). The cumulative set of characters is about twice smaller than that of the *Book of Ndɔlɛ* with some 120 syllabic signs and sixteen ideograms. The majority of the cumulative list is found in the character

repertoire of the *Book of Ndɔlɛ*. Not occurring there are most individual vowels, both oral and nasal (/e, an, o, ɔ, ɛ, ɛn/), and the syllable /ha/. Prenasalized nd-series are written with different shapes in the *Book of Ndɔlɛ*. All the mentioned differences can be explained, most probably, by different orthographic approaches used in the old manuscript and in modern texts (cf. Stewart 1972).

4. Uniformity tests

As we mentioned in the Introduction, previous studies revealed the uniform distribution of complexities for alphabetic scripts but not in the Vai syllabary. The uniformity hypothesis can be tested by the Wald–Wolfowitz runs test (Wald & Wolfowitz 1940; cf. also Stewart 2009, Chap. 17; Rajagopalan 2006, p. 187-188). We demonstrate it here for the cumulative set of characters from Table 2. All the relevant numbers are listed in Table 3 for four Vai texts analyzed in this work. Let $I = 55$ denote the inventory size and $R = 28$ is the range of complexities. The uniform distribution of data means that all expected frequency values equal $E = I/(R+1)$. A run is a sequence of frequencies which are either all greater than E or all smaller than E. For the case under consideration $E \simeq 1.9$ and $r = 12$, namely [1, 2,2,2,2, 1, 3,2,3,3, 1,1,1, 3,2,4, 0, 3,2,3,2,4, 1, 3, 0,1,0,1, 2]. Let $n_1 = 11$ is the number of frequencies smaller than E, $n_2 = 18$ is the number of frequencies larger than E, and $n = R + 1 = 29$.

Table 3
Distribution of complexities in Vai texts.

C \ f_C	OV	UDHR	WWV	PR	Cumulative
5	1	1	—	1	1
6	0	1	1	1	2
7	1	1	1	2	2
8	2	1	1	1	2
9	1	1	2	1	2
10	1	1	1	1	1
11	3	2	0	2	3
12	1	1	0	2	2
13	2	2	2	3	3
14	2	3	1	1	3
15	1	1	1	1	1
16	1	1	1	1	1
17	1	1	1	1	1
18	1	3	2	1	3

C \ f_C	OV	UDHR	WWV	PR	Cumulative
19	1	2	1	1	2
20	3	4	2	4	4
21	0	0	0	0	0
22	2	1	1	3	3
23	2	2	2	2	2
24	3	2	1	2	3
25	2	1	1	2	2
26	2	3	2	4	4
27	1	1	0	1	1
28	1	2	2	2	3
29	0	0	0	0	0
30	1	0	0	0	1
31	1	0	0	0	0
32	1	1	1	1	1
33	1	1	1	1	2
I	38	40	28	42	55
R	28	28	27	28	28
$E = I / (R+1)$	1.3	1.4	1.0	1.5	1.9
n_1	19	19	21	18	11
n_2	10	10	7	11	18
n	29	29	28	29	29
r (runs)	11	13	15	11	12
$E(r)$	14.1	14.1	11.5	14.7	14.7
σ_r	2.38	2.38	1.92	2.48	2.57
z	1.09	0.25	1.56	1.27	1.51

The number of runs is considered random (meaning that the distribution is uniform) if

$$z = |r - E(r)| - 0.5\sigma_r < 1.96,$$

where

$$E(r) = 1 + \frac{2n_1n_2}{n} \quad \text{and} \quad \sigma_r^2 = \frac{2n_1n_2(2n_1n_2 - n)}{n_2(n-1)}.$$

The results of the calculations are presented in Table 3. As one can see, the defined core parts in all the samples, as well as the cumulative list of characters, confirm the uniformity hypothesis.

5. Correlation between complexity and frequency in the Vai syllabary

Previously, the correlation between complexity and frequency was studied for the Nko script (Rovenchak & Vydrin 2010), and the values of the Pearson correlation coefficient $r_P = -0.39$ and the Spearman correlation coefficient $r_S = -0.20$ were obtained. We have analyzed the four Vai texts with respect to this property by calculating the respective correlation coefficients. The full frequency lists were taken into consideration, not only that of the core part. The data are shown in Table 4 in comparison with the Morse code for English.

Table 4
Correlation coefficients in Vai texts

	OV	UDHR	WWV	PR	Nko	Morse code (English)
Pearson	–0.15	–0.09	–0.14	–0.09	–0.39	–0.81
Spearman	–0.20	–0.12	–0.11	–0.13	–0.20	–0.79

Small values of the coefficients suggest that the simplification of shapes is not a prevailing mechanism in the development of a script, but it still has some [marginal] role. Note the values of the correlation coefficients of the Morse code which was artificially created basing on the frequency considerations. Complexity of the Morse code was defined according to the standard duration of its elementary signals, namely 1 for the short signal (‘dot’) and 3 for the long signal (‘dash’).

Interestingly, the highest correlation is obtained for the text being a continuous narrative, namely *Our Village*. Such texts are the most natural for the analysis of frequency structure and for the correlation study in particular.

6. Conclusions

In the paper, we proposed the definition of the *core part* of a syllabary basing on the *h*-point, which is obtainable from the frequency analysis of text. The characters with frequencies higher and equal to the value of the *h*-point are considered as belonging to the core. Previously, it was suggested that some set of the most frequent characters can be tested with respect to the uniformity hypothesis for the distribution of complexities, which otherwise failed for the whole syllabary. This hypothesis was confirmed in all the considered cases for the core part of the Vai syllabary defined as above. The core parts obtained from different texts significantly overlap, and the compiled cumulative list of core characters contains 55 symbols versus 42 symbols in the largest core for an individual text.

No considerable differences of the frequency structure occur between the indigenous texts (OV and VPR) and translated ones (UDHR and WWV). This should facilitate further studies of Vai texts as many of them are translations.

Lastly, we studied the correlation between the complexity of characters and their frequency. Small but negative values of both Pearson's and Spearman's correlation coefficients signal that the simplicity of shapes is not a key feature in the script development. This issue requires a broader analysis, with more texts (in particular, from different periods of time) if speaking specifically about the Vai syllabary and with more scripts in general.

References

Altmann, G. (2004). Script complexity. *Glottometrics 8, 68-74*.

Barrow, John D. (2002). *The constants of nature*. New York: Pantheon Books.

Buk, Solomija (2010). Statystyčna struktura romanu Ivana Franka *Boryslav smijetjsja* [Statistical structure of *Boryslav Laughs*, a novel by Ivan Franko]. *Scientific Notes of Taurida V. I. Vernadsky National University. Series: Philology. Social communications. 23*[*62*](*3*), *114-118.*

Buk, S., Mačutek, J., Rovenchak, A. (2008). Some properties of the Ukrainian writing system. *Glottometrics 16, 63-79.*

Dalby, D. (1967). A survey of the indigenous scripts of Liberia and Sierra Leone: Vai, Mende, Loma, Kpelle and Bassa. *African Language Studies 8, 1–51.*

Daniels, Peter T. (1990). Fundamentals of grammatology. *Journal of the American Oriental Society, 110*(*4*), *727-731*.

E Numbers (2008). In: *The Mudflap: Hertfordshire Wheelers Monthly Newsletter* (December), *2-3.*

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceeding of the National Academy of Sciences USA 102*(*45*), *16569–16572*.

Jones, Adam (1981). Who Were the Vai? *The Journal of African History 22*(*2*), *159-178.*

Lewand, Robert Edward (2000). *Cryptographical mathematics*. Washington, DC: The Mathematical Association of America Publishing.

Mačutek, J. (2008). Runes: complexity and distinctivity. *Glottometrics 16, 1-16.*

Popescu, I.-I. et al. (2009). *Word frequency studies.* Berlin–New York: Mouton de Gruyter.

Rajagopalan, V. (2006). *Selected statistical tests.* New Age International.

Rovenchak, A., Mačutek, J., Riley, C. (2009). Distribution of complexities in the Vai script. *Glottometrics 18, 1–12*.

Rovenchak, A., Vydrin, V. (2010). Quantitative properties of the Nko writing system. In: P. Grzybek, E. Kelih, & J. Mačutek (eds.) *Text and Language:*

Structures - Functions - Interrelations. Quantitative perspectives: 171-181 Wien: Praesens.

Stewart, Gail (1972). The early Vai script found in the Book of Ndole. In: *Conference on Manding Studies 1-27*. London: School of Oriental and African Studies.

Stewart, William J. (2009). *Probability, Markov chains, queues, and simulation: the mathematical basis of performance modelling.* Princeton*:* Princeton University Press.

Wald, A., Wolfowitz, J. (1940). On a test whether two samples are from the same population. *The Annals of Mathematical Statistics 11*(*2*), *147-162*.

Kelih, E. et al. (eds.), Issues in Quantitative Linguistics, Vol. 2, 129-135

Wortlängen im Norwegischen

Karl-Heinz Best

1. Wortlängen im Norwegischen

Wortlängen sind einer der am meisten bearbeiteten Gegenstände im Göttinger *Projekt Quantitative Linguistik* (Best 1998). Die Untersuchungen stützen sich auf die theoretischen Arbeiten von Wimmer u.a. (1994) und Wimmer & Altmann (1996) (vgl. dazu auch Best 2006: 27ff.). Die dort vorgestellte Theorie, dass nämlich die Wahrscheinlichkeit der Wortlängen der Längenklasse x zu der der Längenklasse $x - 1$ proportional ist, hat sich inzwischen bei über 4000 Texten in ca. 50 Sprachen bewährt. Hier geht es darum, einige Daten zum Norwegischen nachzutragen.

2. Datenbasis und Verfahren

Die Darstellung stützt sich auf eine Untersuchung von Jenner (1997), die außer deutschen Texten vor allem norwegische Briefe und Pressetexte auswertete. Die Wortlänge wurde von Jenner danach bestimmt, wie viele Silben je Wort beobachtet wurden; Kriterium für die Silbenzahl ist die Zahl der Vokale je Wort. Es handelt sich bei Jenners Untersuchung um Texte der Bokmål (23 Briefe von Knut Hamsun an seinen Sohn Tore; 20 Briefe des gleichen Autors an seine Frau Marie und 22 Pressetexte, die in *Aftenposten* erschienen sind) sowie des Nynorsk/der Landsmål (20 Pressetexte, die in *Dag og Tid* erschienen). Sie wurden von Jenner (1997) einzeln daraufhin untersucht, ob ihre Wortlängen gemäß der 1-verschobenen Hyperpoisson-Verteilung

$$P_x = \frac{a^{x-1}}{b^{(x-1)}\ {}_1F_1\,(1;b;a)}, \quad x = 1,2,\ldots$$

in ihnen vorkommen. Die Anpassung dieser Verteilung gelang in allen Fällen; lediglich in 2 Fällen (je ein Text der Pressesprache in Bokmål und Nynorsk) war das Ergebnis nicht ganz zufriedenstellend. Die 1-verschobene Hyperpoisson-Verteilung hat sich damit auch für norwegische Texte als ein geeignetes Modell für die Wortlängenverteilungen erwiesen.

In dieser Untersuchung werden nun nicht die Texte einzeln bearbeitet; stattdessen werden die Texte der vier genannten Textgruppen jeweils für sich

zusammengefasst und daraufhin geprüft, mit welchem Ergebnis die gleiche Verteilung an diese neu gebildeten Dateien angepasst werden kann.

3. Ergebnis der Anpassung der 1-verschobenen Hyperpoisson-Verteilung

Die Anpassung der 1-verschobenen Hyperpoisson-Verteilung mit Hilfe des *Altmann-Fitters* (1997) an die Dateien ergab die Resultate in Tabelle 1. Die Briefe stammen aus den Jahren 1915 – 1948.

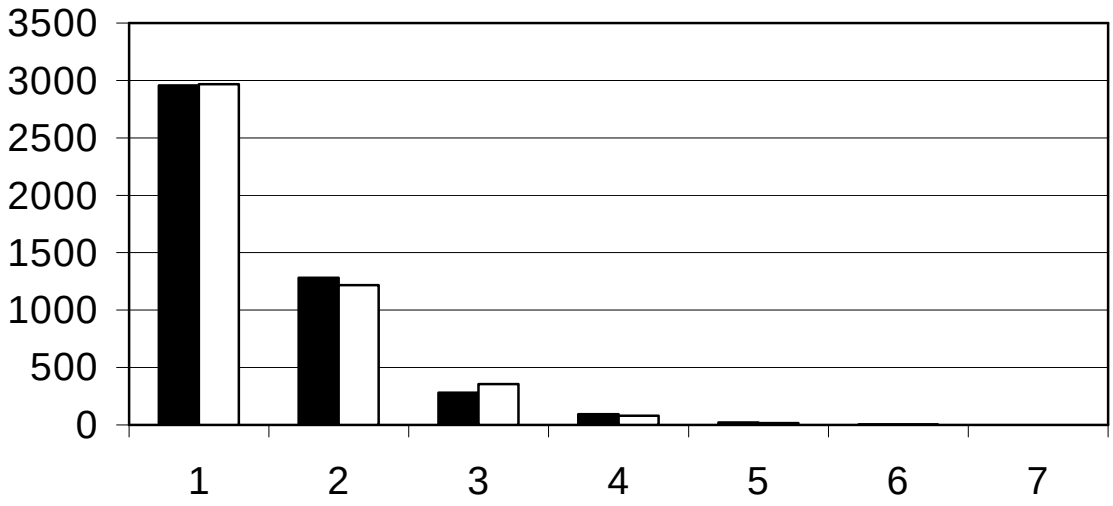
Tabelle 1
Wortlängen in 23 Briefen von Hamsun
an seinen Sohn Tore (Jenner 1997, 67-72)

x	n_x	NP_x	x	n_x	NP_x
1	2958	2966.61	5	21	15.07
2	1282	1216.55	6	2	2.37|
3	280	356.03	7	1	0.37|
4	94	81.00			
a = 1.0220 b = 2.4922, C = 0.01					

Legende zur Tabelle 1 (weitere Erläuterungen in Best 2006: 29-33.):
x: Wortlänge (in Silben);
n_x: beobachtete Zahl der Wörter mit der Silbenzahl x;
NP_x: aufgrund der 1-verschobenen Hyperpoisson-Verteilung errechnete Zahl der Wörter mit der Silbenzahl x;
a, b: Parameter der Verteilung.

C ist der Diskrepanzkoeffizient, χ^2 / N, der bei sehr umfangreichen Dateien eingesetzt wird. Er signalisiert mit $C \leq 0.01$ eine gute Übereinstimmung zwischen Modell und Beobachtung. Diese Bedingung ist hier ebenso wie in den noch folgenden drei Fällen erfüllt.
Die senkrechten Linien in den Tabellen bedeuten, dass die betroffenen Längenklassen bei der Anpassung der 1-verschobenen Hyperpoisson-Verteilung zusammengefasst wurden.

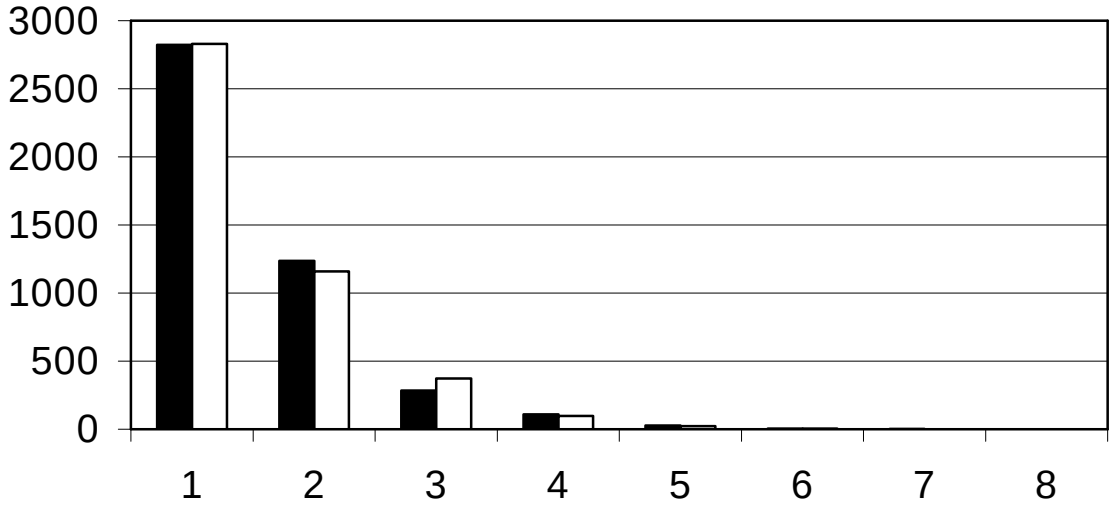
Die graphische Darstellung der Anpassung ist in Graphik 1 zu sehen. Die Wortlängen in Hamsuns Briefen an seine Frau Marie findet man in Tabelle 2 und Graphik 2. Diese Briefe stammen aus den Jahren 1909 – 1932.



Graphik 1: Wortlängen in Briefen von Hamsun an seinen Sohn Tore
(schwarz: beobachtete Werte; weiß: berechnete Werte)

Tabelle 2
Wortlängen in Briefen von Hamsun an seine Frau Marie (Jenner 1997, 72-77)

x	n_x	NP_x	x	n_x	NP_x
1	2823	2830.06	5	28	21.91
2	1237	1159.86	6	4	4.25|
3	283	372.03	7	2	0.73|
4	109	98.02	8	1	0.13|
a = 1.4756, b = 3.6004, C = 0.0067					



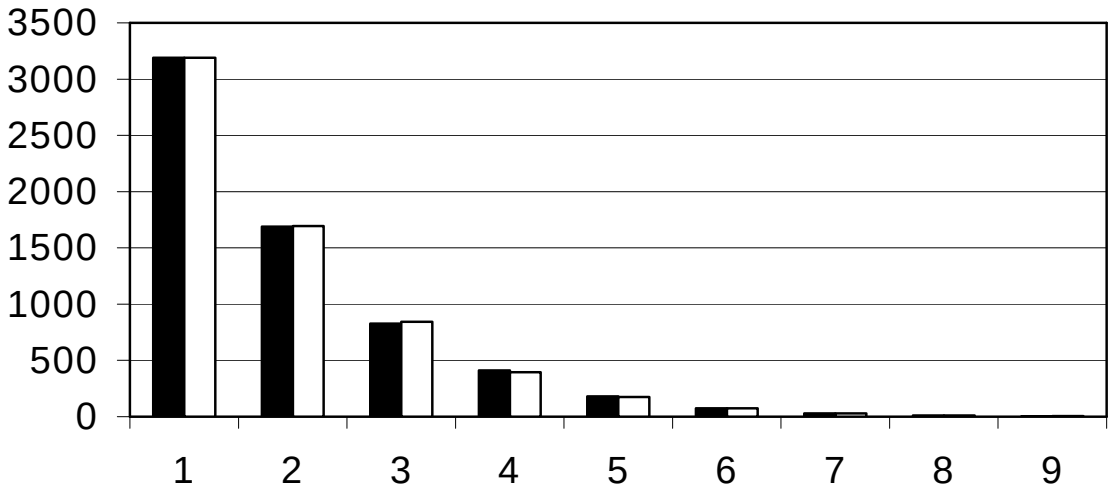
Graphik 2: Wortlängen in Briefen von Hamsun an seine Frau Marie

Zum Vergleich zeigen wir die Verteilung von Wortlängen in norwegischer Presse in Tabelle 3 und 4 mit graphischer Darstellung in den Graphiken 3 und 4.

Tabelle 3
Wortlängen in Pressetexten in *Aftenposten* (Bokmål) (Jenner 1997, 92-99)

x	n_x	NP_x	x	n_x	NP_x
1	3191	3188.70	6	73	72.57
2	1689	1695.40	7	29	28.73
3	827	843.57	8	10	10.82
4	412	394.42	9	2	5.86
5	181	173.93			
a = 7.7530, b = 14.5818, C = 0.0006					

Diese Pressetexte stammen aus den Jahren 1994 und 1996.

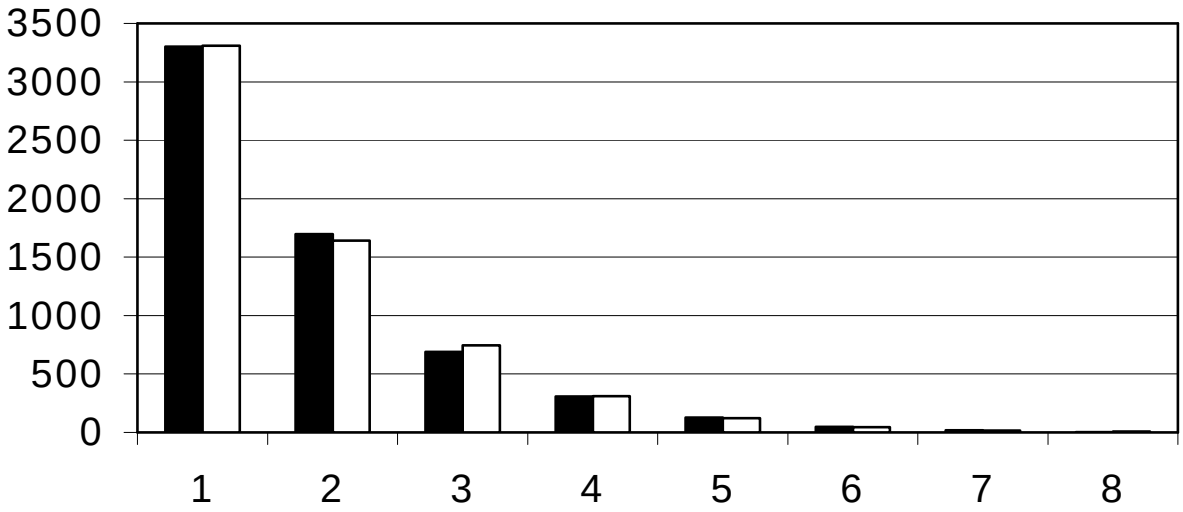


Graphik 3: Wortlängen in Pressetexten in *Aftenposten*

Tabelle 4
Wortlängen in Pressetexten in *Dag og Tid* (Nynorsk) (Jenner 1997, 99-105)

x	n_x	NP_x	x	n_x	NP_x
1	3301	3309.67	5	126	120.66
2	1697	1640.38	6	48	43.63
3	690	744.08	7	18	14.78
4	308	311.13	8	3	6.67
a = 5.3487, b = 10.7916, C = 0.0015					

Diese Pressetexte stammen aus dem Jahr 1996.



Graphik zu dem Textkorpus aus den drei Einzeltexten

4. Exkurs: Mittelwerte der Wortlängen der vier Textgruppen

Tabelle 5 gibt eine Übersicht über die Mittelwerte der Wortlängen.

Tabelle 5
Mittelwerte der Wortlängen

Varietät	Textgruppe	Wörter	Mittelwert
Bokmål	Hamsun: Briefe an Tore	4638	1.48
Bokmål	Hamsun: Briefe an seine Frau Marie	4487	1.51
Bokmål	Pressetexte aus *Aftenposten*	6414	1.92
Nynorsk	Pressetexte aus *Dag ogTid*	6191	1.79

Wie man sieht, ist die Wortlänge in den beiden Textsorten unterschiedlich. Weitere Untersuchungen wären nötig um festzustellen, ob dies durch die Entwicklung des Norwegischen. im zwanzigsten Jahrhundert oder rein durch Textsortenunterschiede entstand.

5. Zusammenfassung

Die Untersuchung von Jenner (1997) hatte als Ergebnis erbracht, dass an alle einzelnen Texte die 1-verschobene Hyperpoisson-Verteilung angepasst werden kann; nur in zwei Fällen ist das Ergebnis nicht ganz befriedigend.

Hier konnte nun gezeigt werden, dass die gleiche Verteilung auch dann ein gutes Ergebnis erbringt, wenn man die Daten der einzelnen Texte der vier Textgruppen jeweils für sich zusammenfasst.

Damit wird die Hypothese von Wimmer & Altmann (1996) sowie von Wimmer u.a. (1996), dass Wortlängen sich gesetzmäßig verhalten, auch durch die Befunde zum Norwegischen gestützt. Das Norwegische unterliegt damit der gleichen Gesetzmäßigkeit wie auch alle anderen nordgermanischen Sprachen (vgl. dazu jetzt Best 2011). Den Darstellungen von Altmann (1988) kann man entnehmen, dass die gleiche Theorie auch für Satzlängen gilt, den anderen Untersuchungen im Göttinger *Projekt Quantitative Linguistik* ist zu entnehmen, dass auch Morphe (Best 2005b), rhythmische Einheiten (Best 2005a) und Silben (Cassier 2001) den gleichen Gesetzmäßigkeiten folgen.

Literatur

Altmann, G. (1988). *Wiederholungen in Texten*. Bochum: Brockmeyer.

Best, K.-H. (1998). Results and Perspectives of the Göttingen Project on Quantitative Linguistics. *Journal of Quantitative Linguistics 5, 155-162.*

Best, K.-H. (2005a). Längen rhythmischer Einheiten. In: Köhler, R., Altmann, G., Piotrowski, R.G. (Hrsg.), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch: 208-214.* Berlin-New York: de Gruyter.

Best, K.-H. (2005b). Morphlänge. In: Köhler, Reinhard, Altmann, Gabriel, Piotrowski, R.G. (Hrsg.), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch: 255-260.* Berlin-NewYork: de Gruyter.

Best, K.-H. (2006). *Quantitative Linguistik: Eine Annäherung.* 3., stark überarbeitete und ergänzte Auflage. Göttingen: Peust & Gutschmidt.

Best, K.-H. (2011). Wortlängen im Dänischen. *Göttinger Beiträge zur Sprachwissenschaft* (Im Druck).

Cassier, F.-U. (2001). Silbenlängen in Meldungen der deutschen Tagespresse. In: Best, K.-H. (Hrsg.), *Häufigkeitsverteilungen in Texten: 33-42*. Göttingen: Peust & Gutschmidt.

Hamsun, T. (1956). *Knut Hamsun som han var. Et utvalg av hans brev.* Oslo: Gyldendal.

Jenner, K. (1997). *Zur Wortkomplexität deutscher und norwegischer Texte*. Staatsexamensarbeit, Göttingen.

Wimmer, G., Altmann, G. (1996). The theory of word length distribution: some results and generalizations. In: Schmidt, P. (Hrsg.), *Glottometrika 15, 112-133*. Trier: Wissenschaftlicher Verlag Trier.

Wimmer, G., Altmann, G, (1999). *Thesaurus of univariate discrete probability distributions*. Essen: Stamm.

Wimmer, G., Köhler, R, Grotjahn, R., Altmann, G. (1994). Towards a theory of word length distribution. *Journal of Quantitative Linguistics 1, 98-106.*

Software

Altmann-fitter (1997). *Iterative Fitting of Probability Distributions.* Lüdenscheid: RAM-Verlag.

Informationen zum Göttinger *Projekt Quantitative Linguistik*: Homepage: http://wwwuser.gwdg.de/~kbest/

Kelih, E. et al. (eds.), Issues in Quantitative Linguistics, Vol. 2, 136-147

Canonical word forms: Menzerath–Altmann law, phonemic length and syllabic length

Ján Mačutek, Andrij Rovenchak

1. Introduction

Canonical word forms (CWFs henceforth) are words in which phonemes are reduced to consonants C and vowels V (e.g., the CWF of the English word "study" is CCVCV). CWFs in Indonesian were investigated by Altmann et al. (2002: 37–54). A summary of quantitative approaches and results on the topic, including a list of literature, was given by Altmann (2005).

Strauss et al. (2008: 2–3, 5–6) formulated several hypotheses on CWFs, of which some are addressed in this paper. Here, we analyze only CWF types, i.e., each CWF is taken into account once, regardless of the frequencies with which particular CWFs occur.

The paper is organized as follows: After the introduction, we describe our data from Indonesian and Ukrainian. In the next section, we demonstrate that the relation between syllabic length of CWFs and the mean phonemic length of syllables in respective CWFs can be modeled by the Menzerath–Altmann law (MA law henceforth; cf. Cramer 2005a), and in addition, we suggest an interpretation of the parameters of the model. Section 4 is dedicated to the relation between phonemic and syllabic length of CWFs. Altmann et al. (2002: 43–46) and Strauss et al. (2008: 2) model the relation by a linear function. We show that although a linear function yields an excellent fit, it is not consistent with the MA law. On the contrary, a well-fitting non-linear model for the relation can be derived from the MA law deductively, thus preserving the interpretation of the parameters used.

2. Data

In this study, two languages (Indonesian and Ukrainian) from different language families are considered in order to test the linguistic laws and properties in a more general fashion and to reduce the risk of arriving at a language specific model.

The Indonesian data from Altmann et al. (2002: 37–54) were taken and reanalyzed. Texts for Ukrainian were taken from the database collected within the project "Constructing a Balanced Ukrainian Text Databank" (see Kelih et al. 2009 for details). We have analyzed texts of several very different genres, namely: blog, drama, scientific paper (in humanities), scientific paper (in physics),

sermon, sport reportage. Text subcorpora consisting of the above mentioned genres and a corpus consisting of all the texts were investigated.

Ukrainian texts were automatically converted from the grapheme to phoneme level using the principles described by Buk et al (2008). For the purposes of our study, a sophisticated conversion itself is not of great importance as the Ukrainian orthography is quite regular and "shallow" (cf. Coulmas 2004: 380). The main peculiarities to be taken into consideration are:

1. graphemes < я, ю, є > represent two phonemes in a syllable-initial position (/ja, ju, jɛ /, respectively);
2. grapheme < ї > always represents two phonemes /ji/ (in some historical orthographies it had a behavior similar to the graphemes from the previous item);
3. grapheme < щ > always represent two phonemes /ʃtʃ/;
4. grapheme < ь > does not represent any phoneme but is used to mark the palatalization of a preceding consonant;
5. some consonant clusters (< стд, стс, стськ, нтст >, etc.) undergo phonetic simplifications; these are not very frequent, however.

The number of syllables in a Ukrainian word is easily determined. It equals the number of vowels (/a, ɛ, ɪ, i, ɔ, u/) due to quite simple vocalism and the absence of diphthongs. Syllables can be easily counted even without converting a word to phonemes; one has just to count the number of graphemes for vowels < а, е, і, и, о, у > and iotified vowels < я, є, ї, ю >.

For the study on syllabic structures, the presence of zero-syllable words in Ukrainian is important. Such words are very frequent as they denote synsemantic parts of speech. The forms without vowels have vocalized counterparts, and the use of either is determined from the considerations of euphony. Antić et al. (2006) joined zero-syllable words with words which precede or follow them. The same approach was applied also to the Ukrainian data under analysis. Zero-syllable words were treated in the following fashion:

1. particles *б* and *ж* **preceded** by a word were joined with this word (the vocalized counterparts are *би* and *же*, respectively);
2. prepositions *в* and *з* and conjunction *й* **followed** by a word were joined with this word (the vocalized counterparts are *у*, *із/зі/зо*, and *і*, respectively).

A rule of thumb for such a treatment is to determine if the respective zero-syllable word can start a sentence or be used in a sentence-final position.

The variety of CWFs is rich in both Indonesian and Ukrainian (556 and 1578 CWF types in the respective corpora; some typical Indonesian CWFs include CVC, CVCV, CVCVC; the most frequently occurring types in Ukrainian are: CV, CVCV and CVCVC).

3. MA law: syllabic length of CWFs and mean phonemic length of syllables

The MA law describes the relation between sizes (e.g., length) of a language construct and its constituents (e.g., words and syllables, clauses and words, etc). It was observed in many areas of linguistics (cf. the summary paper by Cramer 2005a). Its most general form is expressed by the function

$$(1) \qquad y(x) = ax^b e^{cx},$$

with x being a measure of the construct and $y(x)$ a measure of its constituents. According to the law, the measure $y(x)$ of the constituents decreases with the increasing measure x of the construct, possibly with minor local modification (usually for low values of x).

On the level of word or CWF, length is measured as the number of syllables (W_S) yielding the size of a construct. Syllable length is measured as the number of phonemes or graphemes[1] (S_P; $S_P(W_S)$ denotes the mean phonemic length of syllables in words or CWFs with length W_S) yielding the size of the constituents. It is sufficient to use the function

$$(2) \qquad S_P(W_S) = aW_S^{\,b},$$

a special case of (1) for $c = 0$. Relation (2) has been empirically corroborated for several languages (e.g., Turkish by Hřebíček 1995: 19–21, Croatian by Grzybek 1999, Slovene by Grzybek 2000, Serbian by Kelih 2010). The results by Altmann et al. (2002: 46–48) confirm the validity of law (2) also for CWFs in Indonesian (x is the syllabic length of CWFs, $y(x)$ is the mean phonemic length of syllables in CWFs with x syllables).

The interpretation of the parameters a and b in the MA laws was discussed in general by Köhler (1984, 1989) and Cramer (2005b). Kelih (2010) replaced the parameter a with the mean phonemic length of syllables in one-syllable words (which is the same as the mean phonemic length of one-syllable words), i.e.,

$$(3) \qquad a = S_P(1).$$

The goodness of fit of model (2) with the interpreted parameter a remains acceptable.

1 The number of graphemes was used as a measure of syllable length by Kelih (2010) for Serbian, in which there are almost no differences between the numbers of graphemes and phonemes in words.

Since mean phonemic syllable length decreases with increasing syllabic word length, the parameter *b* in function (2) is negative. The consequence is that (2) converges to 0. However, each syllable contains at least one phoneme. Therefore, we apply a modification (of adding a constant, as suggested by Altmann et al. 2002: 47) of formula (2), namely,

$$(4) \qquad S_P(W_S) = aW_S^{\,b} + 1,$$

respecting thus the minimal syllable length.[2] The interpretation (3) of the parameter *a* must be adjusted analogously

$$(5) \qquad a = S_P(1) - 1.$$

Altmann et al. (2002: 46–48) applied function (2) to Indonesian CWFs, with the determination coefficient $R^2 = 0.93$. We fit the function (4) to seven Ukrainian datasets (six subcorpora and the corpus) described in Section 2. However, we take into account only syllabic lengths satisfying both of the following two conditions: 1) the number of syllables is less than or equal to 10 (behavior of constituents sizes in constructs with a greater size is irregular[3], cf. Kelih 2010: 73), 2) at least 5 CWF types with the given length are observed (to guarantee a certain stability of mean syllable length). The Indonesian data were also reanalyzed. We followed the approach of Kelih (2010) and replaced the parameter *a* with (5). The parameter *b* was estimated by iterative procedures using the software program NLREG. The results are presented in Table 1 below, in which S_{Ptheor} denotes values obtained from function (4). The goodness of fit is satisfactory in all eight cases; the determination coefficient is higher than 0.95 for all data.

2 Buk and Rovenchak (2007) applied the model $S_P(W_S) = aW_S^{\,b} + c$, which is a generalization of (4). This function yields a good fit also in the case of zero-syllable words treated as a separate class.

3 One of reasons could be that a ratio of compounds among long words is (significantly) higher than among short words. If the W_S-S_P relation in the compound components is governed by the components lengths (and not by the compound length), validity of the MA law (with respect to the W_S-S_P relation) in compounds is dubious. The influence of compounds could be relatively strong for long CWF types. This hypothesis has not been tested so far.

Table 1
MA law for CWFs in Indonesian (IND) and Ukrainian (UKR)

IND			UKR-blog			UKR-drama			UKR-humanities		
W_S	S_P	S_{Ptheor}	W_S	S_P	S_{Ptheor}	W_S	S_P	S_{Ptheor}	W_S	S_P	S_{Ptheor}
1	3.60	3.60	1	3.55	3.55	1	3.42	3.42	1	3.93	3.93
2	2.75	2.92	2	3.04	2.98	2	2.95	2.93	2	3.32	3.22
3	2.53	2.60	3	2.84	2.71	3	2.78	2.69	3	2.87	2.89
4	2.33	2.41	4	2.52	2.54	4	2.54	2.54	4	2.65	2.69
5	2.24	2.28	5	2.42	2.41	5	2.40	2.43	5	2.51	2.54
6	2.24	2.18	6	2.25	2.32	6	2.31	2.34	6	2.35	2.44
7	2.26	2.10	7	2.19	2.25	7	2.35	2.28	7	2.33	2.35
8	2.10	2.04				8	2.13	2.22	8	2.29	2.28
									9	2.24	2.22
									10	2.28	2.17
a = 2.60 b = -0.440 R^2= 0.9561			a = 2.55 b = -0.366 R^2= 0.9774			a = 2.42 b = -0.328 R^2 = 0.9787			a = 2.93 b = -0.398 R^2 = 0.9885		

UKR-physics			UKR-sermon			UKR-sport			UKR-corpus		
W_S	S_P	S_{Ptheor}	W_S	S_P	S_{Ptheor}	W_S	S_P	S_{Ptheor}	W_S	S_P	S_{Ptheor}
1	3.85	3.85	1	3.62	3.62	1	3.45	3.45	1	3.93	3.93
2	3.01	3.14	2	3.22	3.05	2	3.09	2.86	2	3.28	3.23
3	2.80	2.81	3	2.79	2.77	3	2.70	2.58	3	2.84	2.90
4	2.58	2.61	4	2.64	2.60	4	2.54	2.41	4	2.67	2.69
5	2.47	2.47	5	2.43	2.47	5	2.38	2.29	5	2.53	2.55
6	2.37	2.36	6	2.29	2.38	6	2.27	2.20	6	2.40	2.44
7	2.32	2.28	7	2.28	2.31	7	2.12	2.13	7	2.34	2.36
8	2.26	2.21							8	2.28	2.29
9	2.19	2.16							9	2.25	2.23
									10	2.27	2.17
a = 2.85 b = –0.411 R^2 = 0.9898			a = 2.62 b = –0.357 R^2 = 0.9691			a = 2.45 b = –0.397 R^2 = 0.9662			a = 2.93 b = –0.396 R^2 = 0.9932		

Figure 1 shows the data and the fitted function (4) for the Indonesian data and for the Ukrainian corpus.

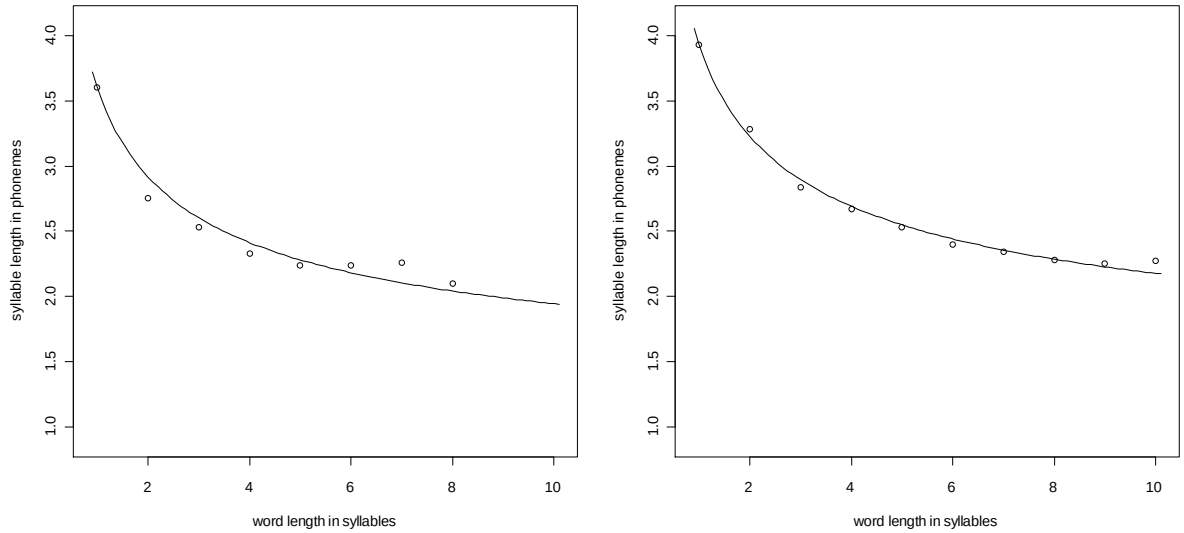


Figure 1. MA law for CWFs in Indonesian (left) and Ukrainian corpus (right).

According to Köhler (1984, 1989), there should be a linear relation between parameters *a* which is interpreted as $S_P(1)$ and *b* of function (2) or (4). Kelih (2010) examined this tendency for the $W_S - S_P$ relation in Serbian ($b = -0.2869S_P(1) + 0.6528$. $R^2 = 0.7109$). Fitting a linear function to the parameters from Table 1, however, does not yield a good fit.[4]

Kelih (2010: 76) then replaced the parameter *b* of function (2) with the corresponding linear function of *a* and obtained

$$(6) \qquad S_P(W_S) = S_P(1)W_S^{-0.2869S_P(1) + 0.6528}$$

The fit of the function (6), which depends on $S_P(1)$ and the two coefficients of the linear function in the exponent, remains satisfactory. On the one hand, the exploitation of the relation between the parameters enables a deeper insight into the "mechanism" of the MA law; on the other hand, however, the parameter *b*, which is at least very generally interpretable as a measure of a shortening tendency (Köhler 1989; Cramer 2005b), is replaced with two uninterpreted coefficients of the linear function.

4 Two possibilities were examined: all parameters (i.e., both Indonesian and Ukrainian ones) and parameters from Ukrainian data only. Neither of them reveals a linear relation. Rather, the values of the two parameters do not seem to be mutually dependent.

4. Relation between phonemic and syllabic length of CWFs

Altmann et al. (2002: 43–46) and Strauss et al. (2008: 2) suggest modeling the relation between CWF length measured in phonemes (W_P) and CWF length measured in syllables by a linear function

$$W_P(W_S) = cW_S + d \,. \quad (7)$$

c and *d* being parameters. $W_P(W_S)$ denotes mean of W_P in CWFs with syllabic length W_S. But for a fixed W_S obviously mean W_P equals mean S_P multiplied by the number of syllables. We thus obtain the equation

$$W_P(W_S) = S_P(W_S) \times W_S \,. \quad (8)$$

Consequently, also the equation

$$S_P(W_S) = \frac{W_P(W_S)}{W_S} \quad (9)$$

is true. The substitution of (7) into (9) yields

$$S_P(W_S) = c + dW_S^{-1}. \quad (10)$$

The exponent of W_S in (10) is fixed to be –1, being thus a special case of the MA law (4). Taking into account the suggested interpretation of the exponent as a measure of a shortening tendency (Köhler 1989; Cramer 2005b), and given that language laws should be general and not language specific, the equation (10) claims that the mean S_P should decrease with increasing W_S at the same rate for all languages and for all text types, which seems to be unrealistic. The values of the exponent can be quite far from –1. Hřebíček (1995: 19–21) obtains b = -0.052 for Turkish; fitting the function (10) to his data yields $R^2 = 0.7894$ which is not very convincing if compared with $R^2 = 0.9307$ for the function (2). Similar values of *b* can be expected especially for other languages in which consonant clusters occur rarely[5]. On the contrary, *b* is a free parameter[6] in either of the forms (2) and (4) of the MA law allowing thus different decrease rates.

5 The optimal value of the parameter *b* is strongly influenced by an additive constant: for the Turkish data, $b = –0.052$ in the model (2), its value is lower ($b = –0.090$) in the model (4), and $b = –0.366$ in the model $S_P(W_S) = aW_S^b + 2$. Consequently, it seems that the (non-)appearance of an additive constant (which itself must be interpreted) in the MA law can play an important role in the interpretation of the parameter *b*.

These theoretical considerations lead us to reject the model (7) tentatively[7] in spite of its excellent fit. The relation between W_P and W_S can, however, easily be derived deductively from the MA law (4) and the equation (8). Substituting (4) into (8) we obtain[8]

$$(11) \quad W_P(W_S) = aW_S^{b+1} + W_S .$$

Since (11) is a corollary of the corroborated linguistic law, the parameter values and interpretations remain the same as in (4), i.e.:

$$(12) \quad W_P(W_S) = \left(S_P(1) - 1\right)W_S^{b+1} + W_S ,$$

with the same values of *b* as in Table 1.

Table 2 contains results of fitting the function (12) to Indonesian and Ukrainian data. The determination coefficient is never less than 0.98.

The data and function (12) fitted for the $W_S - W_P$ relation in Indonesian and in the Ukrainian corpus are presented in Figure 2. One can see that function (12) is very close to a linear function for our parameter values. However, (12) is preferred to (7), because it is theoretically substantiated.

6 Even if a linear relation between the parameters *a* and *b* of the MA law in the form of (2) or (4) is assumed (cf. Köhler 1984, 1989; Kelih 2010), the parameter *b* depends on three factors (the parameter *a* and the two coefficient of the linear function), of which the two latter are free parameters.

7 Nevertheless, function (10) has also an important advantage: it is easy to interpret its parameters. If W_S increases to infinity, the value of $S_P(W_S)$ converges to *c*, which means that *c* is the lower limit of the mean S_P. On the other hand, we have $S_P(1) = c + d$. Thus, the parameter *d* can be interpreted as the difference between the maximum and the lower limit of the mean S_P. We, however, prefer (2) or (4) as the established form of the MA law, unless (11) is corroborated in several typologically different languages.

8 Alternatively, from (2) and (8) it follows that $W_P(W_S) = aW_S^{b+1}$.

Table 2
Fitting function (12) to the $W_S - W_P$ relation
for Indonesian (IND) and Ukrainian (UKR)

IND			UKR-blog			UKR-drama			UKR-humanities		
W_S	W_P	$W_{P\,theor}$	W_S	W_P	$W_{P\,theor}$	W_S	W_P	$W_{P\,theor}$	W_S	W_P	$W_{P\,theor}$
1	3.60	3.60	1	3.55	3.55	1	3.42	3.42	1	3.93	3.93
2	5.50	5.83	2	6.07	5.96	2	5.90	5.86	2	6.65	6.45
3	7.59	7.81	3	8.53	8.12	3	8.33	8.06	3	8.61	8.68
4	9.33	9.65	4	10.07	10.14	4	10.15	10.14	4	10.58	10.75
5	11.21	11.40	5	12.07	12.07	5	11.99	12.14	5	12.57	12.72
6	13.45	13.09	6	13.47	13.94	6	13.84	14.07	6	14.12	14.62
7	15.80	14.73	7	15.36	15.76	7	16.48	15.95	7	16.28	16.45
8	16.82	16.33				8	17.00	17.79	8	18.30	18.25
									9	20.17	20.00
									10	22.78	21.72
	a = 2.60			a = 2.55			a = 2.42			a = 2.93	
	b = -0.440			b = -0.366			b = -0.328			b = -0.398	
	R^2 = 0.9887			R^2 = 0.9946			R^2 = 0.9937			R^2 = 0.9954	

UKR-physics			UKR-sermon			UKR-sport			UKR-corpus		
W_S	W_P	$W_{P\,theor}$	W_S	W_P	$W_{P\,theor}$	W_S	W_P	$W_{P\,theor}$	W_S	W_P	$W_{P\,theor}$
1	3.85	3.85	1	3.62	3.62	1	3.45	3.45	1	3.93	3.93
2	6.03	6.29	2	6.45	6.09	2	6.18	5.72	2	6.57	6.45
3	8.39	8.44	3	8.38	8.31	3	8.11	7.75	3	8.53	8.69
4	10.32	10.45	4	10.56	10.39	4	10.17	9.65	4	10.67	10.77
5	12.33	12.35	5	12.13	12.37	5	11.89	11.47	5	12.66	12.75
6	14.20	14.19	6	13.72	14.29	6	13.62	13.22	6	14.38	14.65
7	16.21	15.97	7	15.94	16.16	7	14.86	14.92	7	16.39	16.49
8	18.08	17.70				8	17.50	16.58	8	18.25	18.29
									9	20.22	20.05
									10	22.71	21.77
	$a = 2.85$			$a = 2.62$			$a = 2.45$			$a = 2.93$	
	$b = -0.411$			$b = -0.357$			$b = -0.397$			$b = -0.396$	
	$R^2 = 0.9983$			$R^2 = 0.9946$			$R^2 = 0.9883$			R^2 = 0.9969	

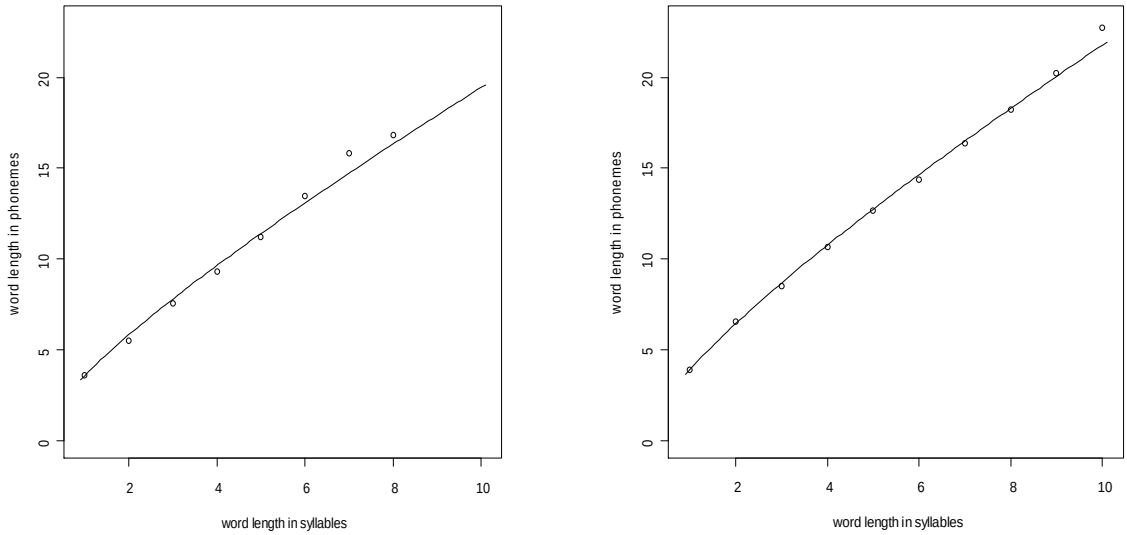


Figure 2. $W_S - W_P$ relation for Indonesian (left) and Ukrainian corpus (right).

5. Conclusion

A systematic relation between syllabic length of canonical word forms and mean phonemic length of syllables was scrutinized. The relation can be modeled by the well-known Menzerath–Altmann law. One of parameters of the model can be interpreted as mean phonemic length of syllables in 1-syllable canonical word forms.

For theoretical reasons, we tentatively reject the hypothesis on the linear relation between syllabic and phonemic length of canonical word forms. A new hypothesis is derived deductively from the Menzerath–Altmann law. Consequently, the parameters of the Menzerath–Altmann law and of the relation between syllabic and phonemic length of canonical word forms have the same values and interpretations. The new hypothesis was empirically corroborated in data from Indonesian and Ukrainian.

Acknowledgement

J. Mačutek was supported by the research grant VEGA 1/0077/09.
The data for Ukrainian were collected within the project M/6-2009 (No. 0109U001786) from the Ministry of Education and Sciences of Ukraine and WTZ Project UA 05/2009 from the Österreichischer Austauschdienst.
The authors wish to thank Eric S. Wheeler for improving the English of the paper.

References

Altmann. G. (2005). Phonic word structure. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 191–208.* Berlin-New York: de Gruyter.

Altmann, G., Bagheri, D., Goebl, H., Köhler, R., Prün, C. (2002). *Einführung in die quantitative Lexikologie.* Göttingen: Peust & Gutschmidt.

Antić, G., Kelih, E., Grzybek, P. (2006). Zero-syllable words in determining word length. In: Grzybek. P. (ed.), *Contributions to the Study of Text and Language. Word Length Studies and Related Issues: 117–156.* Dordrecht: Springer.

Buk, S.. Mačutek, J., Rovenchak, A. (2008). Some properties of the Ukrainian writing system. *Glottometrics 16, 63–79*.

Buk, S., Rovenchak, A. (2007). Statistical parameters of Ivan Franko's novel *Perekhresni stežky (The Cross-Paths).* In: Grzybek, P., Köhler, R. (eds.), *Exact methods in the study of language and text: dedicated to Professor Gabriel Altmann on the occasion of his 75th birthday (Quantitative Linguistics: 62, 39–48.* Berlin-New York: Mouton de Gruyter.

Coulmas, F. (2004). *The Blackwell encyclopedia of writing systems.* Blackwell Publishing.

Cramer, I.M. (2005a). Das Menzerathsche Gesetz. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 659–688.* Berlin-New York: de Gruyter.

Cramer, I.M. (2005b). The parameters of the Menzerath–Altmann law. *Journal of Quantitative Linguistics 12, 41–52*.

Grzybek, P. (1999). Randbemerkungen zur Korrelation von Wort- und Silbenlänge im Kroatischen. In: Tošović, B. (ed.), *Die grammatischen Korrelationen (GraLiS-1999): 67–77.* Graz: Institut für Slawistik der Karl-Franzens-Universität.

Grzybek, P. (2000). Pogostnostna analiza besed iz elektronskega korpusa slovenskih besedil. *Slavistična Revija 48, 141–157.*

Hřebíček, L. (1995). *Text Levels, Language Constructs, Constituents and the Menzerath–Altmann Law*. Trier: WVT.

Kelih, E. (2010). Parameter interpretation of the Menzerath law: evidence from Serbian. In: Grzybek, P., Kelih, E., Mačutek, J. (eds.), *Text and Language. Structures. Functions. Interrelations. Quantitative Perspectives: 71–79.* Wien: Praesens.

Kelih, E., Buk, S., Grzybek, P., Rovenchak, A. (2009). Project Description: Designing and Constructing a Typologically Balanced Ukrainian Text Database. In: Kelih, E., Levickij, V., Altmann, G. (eds.), *Methods of Text Analysis: 125–132*. Chernivtsi: ČNU.

Köhler, R. (1984). Zur Interpretation des Menzerathschen Gesetzes. In: Boy, J., Köhler, R. (eds.), *Glottometrika 6, 177–183*. Bochum: Brockmeyer.

Köhler, R. (1989). Das Menzerathsche Gesetz als Resultat des Sprachverarbeitungsmechanismus. In: Altmann, G., Schwibbe, M. (eds.), *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen: 108–116.* Hildesheim / Zürich / New York: Olms.

Strauss, U., Fan, F., Altmann, G. (2008). *Problems in Quantitative Linguistics 1.* Lüdenscheid: RAM-Verlag.

Kelih, E. et al. (eds.), Issues in Quantitative Linguistics, Vol. 2, 148-169

Semantische Besonderheiten der Komponenten in Komposita mit dem Modell N + N

Yuliya Matskulyak

1. Einleitung

Allgemeingültig ist die Einstellung zur Sprache wie zu einem Zeichensystem, wo den Hauptplatz das Wort einnimmt. In diesem Sinne entsteht die Frage, ob man ein Kompositum wie ein einzelnes Zeichen oder als die Vereinigung zweier einzelner Zeichen betrachten muss (Koeder 1999: 210). In diesem Fall „erfolgt die Sache oder das Objekt, mit dem ein bestimmtes Zeichen verglichen wird, aus der Summe der Merkmale von Komponenten (vielleicht handelt es sich dabei um ein zusammengesetztes Zeichen), sowie deren Verbindungsweise" (Koeder 1999: 221).

Bei der Verbindung zu einem Ganzen von mindestens zwei signifikanten Einheiten, die der Bildung einer neuen Einheit zugrunde liegt, wird der formelle Wortzuwachs mit (von) dem semantischen begleitet (Murjassov 1980). Gerade in diesem Aspekt entstehen mehrere Schwierigkeiten, die vor allem damit verbunden sind, dass die Wortzusammensetzung kein einfaches Addieren von freien Morphemen ist. Denn nach einer ziemlich geringen Musteranzahl gebildeten syntaktischen Strukturen entsprechen sehr reiche und dynamische semantische Beziehungen (Godglück 1997: 21). Für den richtigen Gebrauch und für das Verständnis einer zusammengesetzten Einheit muss der Sprecher mit den semantischen Beziehungen vertraut sein, die formal nicht ausgedrückt sind, aber der Kompositumsstruktur tief zugrunde liegen. Laut den neuesten Forschungen sind „semantische Beziehungen in den Zusammensetzungen in der modernen deutschen Sprache sehr vielfältig" (Braun 1998; Meyer 1993; Motsch 1999; Vandermeeren 1999).

Man soll dabei auch einen Unterschied zwischen einem Kompositum und einer Wortverbindung beachten, denn im ersten Fall ist die Bedeutung eines Kompositums ein abgeschlossenes Ganzes, wogegen in einer Wortverbindung zwei oder mehr Begriffseinheiten nur im Moment des Sprechens eine Verbindung aufnehmen (Henzen 1957).

Die Bedeutung eines Kompositums entspricht nach P. Godglück selten einem lexikalisierten Stand und hängt vom Folgenden ab:

- von der lexikalisierten Bedeutung der in einem Kompositum zusammengesetzten Simplizia oder ihrer metaphorischen, metonymischen oder sonstwie aus dem Standardlexem transformierten Varianten (Simplexbedeutung);
- von der Stellung des Kompositums in der Kohärenzstruktur, der Thema/Rhemaverteilung und der Besetzung semantischer Rollen des Textes, in dem es vorkommt (Textverflechtung);

- von den Weltausschnitten, auf die das Kompositum bzw. seine Teile referieren, samt den ontologischen Bedingungen, unter denen dieser Referenzakt gelingt (Referenzbeziehung);

- von der kognitiven Repräsentation enzyklopädischen oder sprachlichen Wissens, das dem Text, in dem es vorkommt, zugrunde liegt (Wissenspräsentation) (Godglück 1997).

Interessant erscheinen in diesem Sinne auch die Theorien von G. Murphy, die vor allem die Verhältnisse einfacher Konzepte bei ihrer Bildung von komplexen Konzepten betreffen. Es wird dabei behauptet (Murphy 1988, Springer, Murphy 1992), dass die Interpretationstheorien der komplexen Konzepte auch auf die Komposita übertragen werden können.

Die erste Theorie beruht auf der Annahme, dass das komplexe Konzept das Zentrum von beiden Ausgangskonzepten bildet. Die zweite sieht eine asymmetrische Struktur von komplexen Konzepten vor: „Das letzte Konzept in einem Kompositum […] ist das Hauptkonzept, und das erste modifiziert es […]“ (Murphy 1988: 532). Hier geht man davon heraus, dass das einfache Konzept durch eine strukturierte Liste freier Positionen repräsentiert wird, die mit bestimmten Einheiten ausgefüllt werden. Das dritte Modell betrifft vor allem die Kombinationen Adjektiv-Substantiv, in denen dem Adjektiv ein bestimmtes oder ein einziges Merkmal zugeteilt wird, und folglich gleich alle Substantive beeinflusst, mit denen es Verbindung aufnimmt.

Alle aufgezählten Modelle haben sowohl starke, als auch schwache Seiten. Das erste Modell betrifft vor allem so genannte *Ad-hoc*-Komposita (die extra für einen konkreten Fall gebildet werden). Das dritte Modell hat nur ein begrenztes Einsatzgebiet. Auf Grund des zweiten könnte man die große Variationsbreite der Begriffsrepräsentierungen erklären, jedoch kritisieren einige Wissenschaftler „die traditionelle Unterordnungsidee“ (Ungerer, Schmid 1996: 88), die dieser Theorie zugrunde liegt. Sie weist den Zusammensetzungen die Bezeichnung der Begriffe von Unterkategorien zu, die die Merkmale des mit dem Hauptwort genannten Kategorienkonzeptes haben, dem sich danach die Merkmale des mit dem Bestimmungswort (Substantiv oder Adjektiv) genannten Konzeptes anschließen.

Gleichzeitig geben die meisten Wissenschaftler zu, dass sich die Bedeutung neuer Wörter nur unter der Einbeziehung des so genannten Allgemeinwissens erschließen lässt (Tafreschi 2006). Eine große Rolle spielen dabei die Kenntnisse des Sprechers vom genannten Objekt. Diese Meinung teilt auch J. Taylor und betont dabei, dass am Beispiel der Komposita die wichtige Rolle des allgemeinen Wissens deutlich wird und es gerade davon abhängt, ob die Wortkombination interpretiert werden kann (Taylor 2003).

Nach S. Langer lässt sich fast jedes deutsche Substantiv mit einem anderen zusammenschließen und ein Kompositum bilden (Langer 1998). Trotz vieler Ideen und Ansichten bezüglich semantischer Besonderheiten von Zusammensetzungen bleibt die Tatsache unbestreitbar, dass für die Feststellung der Bedeutung eines Kompositums die Bestimmung der Semantik seiner Bestandteile nötig ist.

2. Zielsetzung und Materialien

Das Ziel unserer Untersuchung ist das Erschließen der semantischen Besonderheiten der Substantive als Bestimmungs- und Hauptwörter in einem Kompositum sowie ihr stilistischer und funktionaler Vergleich.

Die Semantik der Komponenten eines Kompositums kann man entweder auf der Wortebene oder auf der Wortklassenebene erforschen. In unserer Arbeit behandeln wir gerade diese Einheiten – lexikalisch-semantische Wortklassen. Sie werden charakterisiert als „lexikalische Felder der paradigmatischer Art, die komplexe Gruppierungen bilden und deren Glieder eine gemeinsame Bedeutung haben […]“ (Vassiljev 1990: 126).

Als das Material dienen 7 406 Substantivkomposita mit dem Modell Nomen + Nomen (weiter N + N). Diese Einheiten wurden aus den belletristischen, publizistischen und wissenschaftlichen Texten extrahiert und schließen insgesamt 34 lexikalisch-semantische Unterklassen (LSU) der Substantive ein (siehe dazu Lewizkij, Matskulyak 2009).

3. Semantische Unterklassen von Substantiv-Bestimmungswörtern

Zu den Haupteigenschaften der LSU, nach denen sie sich vergleichen lassen, gehören ihr Umfang, ihre Gebrauchsfrequenz und ihre Wortbildungsaktivität. Der Umfang einer LSU schließt alle festgestellten Einheiten mit der entsprechenden Semantik ein. Die Gebrauchsfrequenz zeigt uns, wie oft diese Einheiten im Text vorkommen, also die Zahl der Belege.

Um die durchschnittliche Aktivität von Substantiv-Bestimmungswörtern zu bestimmen, haben wir die Formel des arithmetischen Mittels eingesetzt (Perebyjnis 2002: 35):

$$\overline{X} = \frac{\sum x_i n_i}{\sum n_i},$$

wo x_i – der Variablenwert, n_i – die Frequenz ist.

Die Gesamtzahl der Substantive (7 406) und die Zahl der Lexeme (2 540), die als Bestimmungswörter fixiert wurden, ergeben uns die durchschnittliche Aktivität von Substantiv-Bestimmungswörtern – 3 Belege. Nach diesen Angaben können wir alle LSU mit einer höheren Wortbildungsaktivität als hochaktiv und diejenigen mit der niedrigeren als inaktiv bezeichnen.

Um die hochfrequenten Unterklassen zu bestimmen, haben wir die Gesamtzahl der Belege durch die Zahl der LSU dividiert und erhielten 7406/34 = 218 Einheiten. Also halten wir die LSU mit der größeren Gebrauchsfrequenz für hochfrequent.

Die Angaben zum Umfang, Gebrauchsfrequenz und Wortbildungsaktivität der LSU von Substantiv-Bestimmungswörtern führen wir in der Tabelle 1 an.

Tabelle 1
LSU-Verteilung von Substantiv-Bestimmungswörtern im Model N + N

№	LSU	Umfang	Gebrauchs-frequenz	Wort-bildungs-aktivität	Die häufigsten Einheiten
1	Person	273	604	2,5	*Mensch* (25), *Frau* (19)
2	Tiere	84	205	2,4	*Schweine* (14), *Pferd* (12)
3	Somatismen	94	333	3,5	*Hand* (23), *Kopf* (19)
4	Attribute des Menschen	29	67	2,3	*Schuh* (6), *Hose* (5)
5	Pflanzen	118	317	2,7	*Kartoffel* (19), *Pflanze* (17)
6	Stoffe und Materialien	149	489	3,3	*Wasser* (38), *Holz* (30)
7	Raum und Ort	114	521	4,6	*Land* (48), *Stadt* (39)
8	Gebäude und Bauten	96	305	3,2	*Haus* (35), *Küche* (23)
9	Gegenstände und Instrumente	231	519	2,2	*Waffe* (14), *Buch* (12)
10	Essen und Getränke	56	162	2,9	*Bier* (13), *Milch* (11)
11	Anzahl, Maßeinheiten	47	119	2,5	*Atom* (12), *Teil* (10)
12	Bewegung	51	138	2,7	*Reise* (20), *Verkehr* (18)
13	Tätigkeit, Aktion	121	420	3,5	*Arbeit* (61), *Regierung* (22)
14	Dasein	44	173	3,9	*Leben* (43), *Tod* (13)
15	Possessorische Sphäre	18	57	3,2	*Gut* (12), *Besitz* (5)
16	Mentale Sphäre	36	107	3	*Geist* (11), *Traum* (9)
17	Wahrnehmung	14	28	2	*Sinn* (9), *Aussicht* (3)
18	Seelische Sphäre	43	130	3	*Wahl* (33), *Liebe* (9)
19	Sprache und Rede	82	249	3	*Wörter* (25), *Sprache* (18)
20	Physiologische Sphäre	30	75	2,5	*Hunger* (14), *Atem* (7)
21	Verhalten und Handlungen	162	373	2,3	*Krieg* (30), *Kampf* (15)
22	Eigenschaften des Menschen	15	39	2,6	*Energie* (15), *Kraft* (7)
23	Naturerscheinungen und Zustände	42	158	3,8	*Natur* (17), *Dampf* (14)
24	Physikalische Eigenschaften	32	91	2,8	*Wärme* (16), *Licht* (11)
25	Zeit, Alter	40	219	5,5	*Zeit* (38), *Jahr* (20)
26	Kennwerte und Eigen-schaften der Gegenstände	78	200	2,6	*Seite* (16), *Verfahren* (13)
27	Veranstaltung, Spiel	42	96	2,3	*Markt* (12), *Sport* (8)

28	Eigennamen	93	131	1,4	*Europa* (6), *Bosnien* (4)
29	Staat, seine Attribute	31	211	6,8	*Staat* (41), *Wirtschaft* (24)
30	Dokumente, Geld	32	102	3,2	*Steuer* (15), *Finanz* (12)
31	Termini	88	131	1,5	*Kristallisation* (7), *Absorption* (6)
32	Sammelbezeichnungen von Menschen, Organisationen	68	379	5,6	*Volk* (37), *Partei* (34)
33	Abstrakte Begriffe	43	144	3,3	*Macht* (21), *Sicherheit* (19)
34	Wissenschaft, Kultur, Traditionen	44	114	2,6	*Kultur* (24), *Kunst* (14)

Wie Tabelle 1 zeigt, werden als Bestimmungswörter Personenbezeichnungen, z.B. *Brautpaar, Kommunistenherz, Führerweisung, Menschenmasse* (604 Belege), Raum- und Ortsbezeichnungen, z.B. *Planetenberg, Bahngelände, Straßenschild, Landesbischof* (521), Gegenstände- und Instrumentenbezeichnungen, z.B. *Klammerinstrument, Patronenhülse, Löffelreihe, Bettjungfer, Waffenschmied* (519), Stoffe und Materialien, z.B. *Kupferzeit, Kristallart, Sandstraße, Wasserkanister* (489), Tätigkeits- und Aktionsbezeichnungen, z.B. *Lenkungsfunktion, Funktionskampf, Jagdwurst, Baugeschäft, Arbeitsweg* (420), Sammelbezeichnungen von Menschen, Organisationen, z.B. *Delegationschef, Bündnispolitik, Parteifreund* (379), Verhaltens- und Handlungsbezeichnungen, z.B. *Abgabenquote, Regelungssystem, Kontaktschuld, Prozeßrechner, Auftragskatalog, Hilfsarbeiter* (373), Somatismen, z.B. *Beinschelle, Leberfleck, Zungenschleck, Zahnarzt, Blutreizker, Handkarre* (333), Pflanzennamen, z.B. *Anisegerling, Holunderbeere, Kiefernharz, Blumenausfahrer, Fruchtschuppe* (317), Gebäude- und Bautenbezeichnungen, z.B. *Lazarettsaal, Kaminflamme, Zimmerdecke, Kolonnenvolumen, Küchenherd* (305) am häufigsten gebraucht.

Sehr selten werden in dieser Funktion die Einheiten der LSU „Wahrnehmung“, z.B. *Besichtigungsprogramm, Genußmittel, Sinneswandel* (28 Belege), der LSU „Eigenschaften des Menschen“, z.B. *Empfindlichkeitsbereich, Freundlichkeitsstrahlung, Verantwortungs-Imperialismus* (39), der LSU „Possessorische Sphäre“, z.B. *Anschaffungspreis, Verlustliste, Immobilienmakler, Einkommensschicht* (57), der LSU „Attribute des Menschen“, z.B. *Stiefelhose, Hutkappe, Kragenweite, Rüstungskram* (67), der LSU „Physiologische Sphäre“, z.B. *Ermüdungserscheinung, Migränetheorie, Hungermahl* (75) fixiert. Auf solche Wiese können wir zusammenfassen, dass als Bestimmungswörter gewöhnlich jene Einheiten vorkommen, die konkrete Begriffe bezeichnen, und viel seltener diejenigen, die Abstraktionen nennen.

Die höchste Wortbildungsaktivität (die Fähigkeit zur Bildung der Wortbildungsreihen) kennzeichnet:

- die Einheiten der LSU „Staat, seine Attribute“, z.B. *Apartheidsünde, Heerespanzer, Gemeindevorsteher, Militärdiktatur, Wirtschaftspolitik* (6,8

Belege). Hier treffen wir auch eine der längsten Wortbildungsreihen mit dem Substantiv *Staat*:

Staat(s)- *-angehörigkeit, -angestellte, -anwalt, -apparat, -beamter, -besuch, -bibliothek, -bürger, -chef, -diener, -domäne, -duma, -form, -führung, -funktionär, -gefangener, -geschäft, -gewalt, -kanzlei, -kleid, -kommissar, -kosten, -lehre, -limousine, -macht, -melodie, -minister, -partei, -polizei, -präsident, -rat, -recht, -schulden, -sekretär, -sekretariat, -sektor, -sicherheit, -verband, -verfassung, -verleumdung, -zeremonie*

- Sammelbezeichnungen von Menschen und Organisationen, z.B. *Delegationschef, Behördentext, Bündnissolidarität, Gewerkschaftspartei* (5,6 Belege), mit der längsten Wortbildungsreihen zu *Volk*:

Volk(s)/ Völker- *-ball, -bund, -freundschaft, -schlacht, -wanderung, -abstimmung, -aufmarsch, -begehren, -bildung, -deutsche, -dichter, -entscheid, -feind, -gemeinschaft, -genosse, -haus, -held, -küche, -kunst, -lied, -masse, -nationalismus, -partei, -polizist, -rächer, -redner, -republik, -stamm, -sturm, -tanz, -union, -wagen, -wahl, -wirtschaft, -zählung, -zeitung, -zorn*

- Zeit- und Altersbezeichnungen, z.B. *Ferienwohnung, Zukunftsentwurf, Jugendsünde, Morgengrauen* (5,5 Belege), mit der längsten Wortbildungsreihen zu *Zeit*:

Zeit(en)- *-abschnitt, -akkord, -alter, -aufnahme, -bedarf, -budget, -dauer, -diagnostik, -einheit, -loch, -ermittlung, -ersparnis, -faktor, -gefühl, -geist, -genosse, -geschehen, -geschichte, -historiker, -lauf, -lohn, -lupe, -marke, -maß, -minderung, -mode, -not, -plan, -planung, -punkt, -raum, -schätzung, -schrift, -situation, -spanne, -zähler, -zeuge, -zone*

- Raum- und Ortsbezeichnungen, z.B. *Strandkombination, Erdball, Gartenweg, Weltuntergang* (4,6), mit der längsten Wortbildungsreihen zu *Land*:

Land(es)/ Länder- *-adel, -arbeiter, -besitz, -enteignung, -dieb, -parlament, -polizei, -ausschuss, -bedienstete, -bibliothek, -bischof, -brauch, -chef, -fürst, -geschichte, -liste, -meisterschaft, -politik, -regierung, -sitte, -sprache, -teil, -verband, -verräter, -vorsitzende, -vorstand, -währung, -fläche, -flucht, -gebiet, -gericht, -haus, -karte, -kreis, -pflanzen, -plage, -rat, -bruder, -leute, -mädchen, -mann, -straße, -strich, -tag, -vergabe, -volk, -wehr, -wirtschaft*

- Daseinsbezeichnungen, z.B. *Abenteuerliteratur, Entstehungsort, Lebensende* (3,9), mit der längsten Wortbildungsreihen zu *Leben*:

Leben(s)- -art, -dauer, -ende, -erweiterung, -form, -formel, -frage, -fülle, -gefahr, -gefühl, -gemeinschaft, -geschichte, -gewohnheit, -gier, -glück, -hilfe, -jahr, -lage, -lauf, -leistung, -licht, -lust, -mittel, -möglichkeit, -mut, -philosophie, -qualität, -raum, -standard, -stil, -tag, -tätigkeit, -teil, -umstände, -unterhalt, -verfassung, -verurteilter, -wandel, -wechsel, -weise, -werk, -zeichen, -zeit

- Namen der Naturerscheinungen und Zustände, z.B. *Katastrophenspiel, Brandmauer, Feuerpatsche, Dampfdurchtritt* (3,8), mit der längsten Wortbildungsreihen zu *Natur*:

Natur- -beobachtung, -darm, -ereignis, -erkenntnis, -geschichte, -gewalt, -kautschuk, -park, -prozeß, -recht, -seide, -stoff, -talent, -theater, -treppe, -vorkommen, -wissenschaft

- Somatismen, z.B. *Beinschelle, Zehenknochen, Stirnpanzerung, Augenlid, Handrücken* (3,5), mit der längsten Wortbildungsreihen zu *Hand*:

Hand- -arbeit, -ball, -ballen, -bewegung, -fesseln, -fläche, -geld, -gelenk, -gepäck, -granate, -griff, -karre, -koffer, -linie, -rücken, -schelle, -schrift, -schuh, -tasche, -teller, -tuch, -wagen, -wurzel

- Tätigkeits- und Aktionsbezeichnungen, z.B. *Explosionsdruck, Abbaureaktion, Bruchstück, Produktionsfaktor* (3,5), mit der längsten Wortbildungsreihen zu *Arbeit*:

Arbeit(s)- -ablauf, -angebot, -anzug, -aufwand, -bedingung, -beginn, -begriff, -belastung, -bereich, -dienst, -druck, -einsatz, -fluß, -frage, -fülle, -gang, -gasse, -gebiet, -gemeinschaft, -geräusch, -gesellschaft, -hemd, -jahr, -kampf, -kleidung, -kraft, -lärm, -leistung, -lücke, -lust, -maid, -markt, -maschine, -methode, -mittel, -möglichkeit, -niederlegung, -organisation, -plan, -planung, -platz, -produktivität, -puls, -raum, -schluß, -stelle, -stunde, -tag, -teilung, -temperatur, -umsatz, -verfahren, -vermittlung, -volk, -vorgang, -weg, -welt, -zeit, -zimmer, -zuteilung, -zwang

Im Gegensatz dazu ist für Eigennamen (1,4) und Termini (1,5) Bildung von Wortbildungsreihen nicht charakteristisch.

Diese Daten erlauben uns zwei Kategorien der häufig gebrauchten semantischen Unterklassen zu unterscheiden:

1. Diejenigen, die dank der Vielfältigkeit und Fülle ihrer lexikalischen Einheiten häufig gebraucht werden. Darunter subsummieren wir die LSU „Person", die LSU „Gegenstände und Instrumente", die LSU „Verhalten und Handlungen", die LSU „Pflanzen" und die LSU „Sprache und Rede".

2. Diejenigen, die dank der hohen Wortbildungsaktivität ihrer Einheiten häufig gebraucht werden. Darunter befinden sich die LSU „Raum und Ort", die LSU „Stoffe und Materialien", die LSU „Tätigkeit, Aktion", die LSU „Sammelbezeichnungen von Menschen, Organisationen", die LSU „Somatismen", die LSU „Gebäude und Bauten" und die LSU „Zeit, Alter".

Beide Kategorien sind fast gleich nach der LSU-Anzahl und schließen sowohl konkrete, als auch abstrakte Begriffe, sowohl Lebewesen-, als auch Gegenstandsbezeichnungen ein.

Betrachten wir den Gebrauch von Unterklassen der Substantiv-Bestimmungswörter gesondert für jeden Stil (siehe Tab. 2 und Abb. 1), so sehen wir eine gewisse Differenz in der Verteilung der LSU. Termini werden z.B. im wissenschaftlichen Stil (Rang 3) viel häufiger als in der Publizistik (Rang 25) und Belletristik (Rang 31) gebraucht; Tiernamen kommen vor allem in der Belletristik (Rang 10) vor, wobei in der Publizistik und im wissenschaftlichen Stil sie fast ungebräuchlich sind (Rang 33,5 und 32); dasselbe gilt für Somatismen (Ränge: 5-20,5-26,5); Pflanzennamen werden nur selten in der Publizistik gebraucht (Rang 29), stattdessen treffen wir sie häufig im wissenschaftlichen Stil (Rang 6) und in der Belletristik (Rang 7); Personenbezeichnungen sind für die Belletristik (Rang 1) und Publizistik (Rang 2) typisch, viel weniger aber für den wissenschaftlichen Stil (Rang 21).

Zugleich ist der Gebrauch von etlichen Unterklassen ziemlich ähnlich. Das betrifft vor allem Bezeichnungen der Bewegungen (Ränge: 18-19-19); Begriffe der physiologischen Sphäre (Ränge: 27-25-29); Raum- und Ortsbezeichnungen (Ränge: 3-6-5); Veranstaltungen und Spiele werden fast gleich oft in der Belletristik und in der Publizistik (Ränge: 22,5-22) genannt und nur etwas weniger im wissenschaftlichen Stil (Rang 26,6); die LSU „Wissenschaft, Kultur, Traditionen" (Ränge: 24-20,5-22,5) sowie die LSU „Wahrnehmung" (Ränge: 34-31,5-30) zeigen nur geringe Unterschiede für alle drei Stile.

Tabelle 2
Gebrauchsfrequenz und Ränge der LSU von Substantiv-Bestimmungswörtern im Modell N + N in verschiedenen Stilen

	LSU	Belletristik	Rang	Publizistik	Rang	wissenschaftlicher Stil	Rang
1	Person	480	1	111	2	31	21
2	Tiere	194	10	3	33,5	8	32
3	Somatismen	305	5	22	20,5	14	26,5
4	Attribute des Menschen	61	25,5	5	31,5	1	34
5	Pflanzen	237	7	7	29	76	6
6	Stoffe und Materialien	331	4	37	14	130	2
7	Raum und Ort	394	3	78	6	81	5
8	Gebäude und Bauten	252	6	28	18	32	20

9	Gegenstände und Instrumente	405	2	57	7	72	7
10	Essen und Getränke	135	14	3	33,5	25	24
11	Anzahl, Maßeinheiten	51	28,5	33	15	44	11
12	Bewegung	88	18	23	19	33	19
13	Tätigkeit, Aktion	166	12	85	5	192	1
14	Dasein	102	16	43	11	41	12,5
15	Possessorische Sphäre	17	33	12	25	28	22,5
16	Mentale Sphäre	61	25,5	13	23	35	18
17	Wahrnehmung	15	34	5	31,5	10	30
18	Seelische Sphäre	74	21	49	9	9	31
19	Sprache und Rede	147	13	50	8	63	9
20	Physiologische Sphäre	53	27	12	25	11	29
21	Verhalten und Handlungen	219	9	104	3	68	8
22	Eigenschaften des Menschen	18	32	8	28	17	25
23	Naturerscheinungen und Zustände	109	15	10	27	40	14
24	Physikalische Eigenschaften	48	30	6	30	38	15
25	Zeit, Alter	168	11	32	16	36	16,5
26	Kennwerte und Eigenschaften der Gegenstände	82	20	38	13	83	4
27	Veranstaltung, Spiel	71	22,5	16	22	14	26,5
28	Eigennamen	95	17	29	17	7	33
29	Staat, seine Attribute	84	19	101	4	36	16,5
30	Dokumente, Geld	51	28,5	44	10	12	28
31	Termini	25	31	12	25	95	3
32	Sammelbezeichnungen von Menschen, Organisationen	232	8	118	1	52	10
33	Abstrakte Begriffe	71	22,5	40	12	41	12,5
34	Wissenschaft, Kultur, Traditionen	65	24	22	20,5	28	22,5

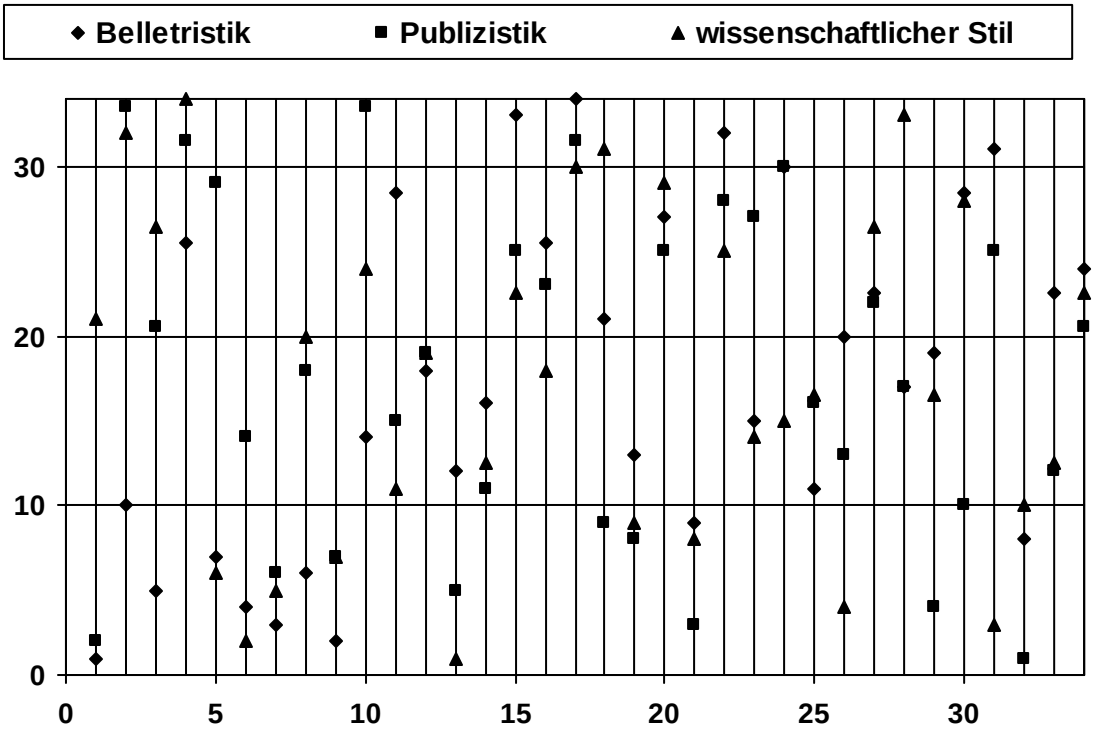


Abbildung 1. Rangkorrelation der LSU von Substantiv- Bestimmungswörtern im Modell N + N in verschiedenen Stilen

Die Korrelationsanalyse, die den Zusammenhang und die Abhängigkeit zwischen verschiedenen Eigenschaften bestimmen lässt und mithilfe der Formel $r_s = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)}$ berechnet wird (d – Rängendifferenz, n – Zahl der korrelierenden Paare), zeigt uns effiziente Ähnlichkeitswerte für den Gebrauch von semantischen Unterklassen der Substantiv-Bestimmungswörter in verschiedenen Stilen (siehe Tab. 3). Das gilt für alle drei Paare bei P = 0,05, wo die Werte höher als 0,35 sind. Bei P = 0,01 (Mindestwerte 0,45) wird das Paar „Belletristik-wissenschaftlicher Stil“ ausgeschlossen. So kommen wir zum Schluss, dass in der Funktion von Substantiv-Bestimmungswörtern belletristische und publizistische Texte am Nächsten sind.

Tabelle 3
Ähnlichkeit der Stile im Gebrauch der LSU
von Substantiv-Bestimmungswörtern im Modell N + N

	Belletristik	Publizistik	Wissenschaftlicher Stil
Belletristik		0,461	0,355
Publizistik			0,453
Wissenschaftlicher Stil			

4. Semantische Unterklassen der Substantiv-Hauptwörter

Unter den Hauptwörtern im Modell N + N haben wir ebenfalls alle semantischen Unterklassen der Substantive fixiert. Auch für diese Kategorie beträgt die durchschnittliche Gebrauchsfrequenz einer LSU 218 Belege und die Wortbildungsaktivität gleicht 3. Dementsprechend können wir bestimmen, für welche semantischen Unterklassen der Hauptwörter die Tendenz zum hohen Gebrauch im Text und hoher Aktivität der Lexeme kennzeichnend ist. Die Angaben zum Umfang, zur Gebrauchsfrequenz und Wortbildungsaktivität zeigt die Tabelle 4.

Tabelle 4
LSU-Verteilung von Substantiv-Hauptwörtern im Model N + N

№	LSU	Umfang	Gebrauchs-frequenz	Wort-bildungs-aktivität	Die häufigsten Einheiten
1	Person	281	689	2,5	*Führer* (23), *Mann* (20)
2	Tiere	42	69	1,6	*Hund* (5), *Käfer* (5)
3	Somatismen	91	234	2,6	*Kopf* (22), *Gesicht* (12)
4	Attribute des Menschen	50	120	2,4	*Mantel* (9), *Kleid* (8)
5	Pflanzen	71	209	2,9	*Baum* (17), *Blatt* (15)
6	Stoffe und Materialien	67	196	2,9	*Stoff* (15), *Stein* (14)
7	Raum und Ort	114	564	4,9	*Weg* (32), *Platz* (24)
8	Gebäude und Bauten	98	404	4,1	*Haus* (41), *Raum* (27)
9	Gegenstände und Instrumente	291	805	2,8	*Bild* (25), *Wagen* (24)
10	Essen und Getränke	62	129	2,1	*Suppe* (13), *Brot* (8)
11	Anzahl, Maßeinheiten	67	248	3,7	*Stück* (25), *Teil* (22)
12	Bewegung	58	164	2,8	*Bewegung* (21), *Fall* (17)
13	Tätigkeit, Aktion	123	320	2,6	*Anlage* (18), *Schlag* (14)
14	Dasein	56	181	3,2	*Geschichte* (26), *Wechsel* (17)
15	Possessorische Sphäre	21	60	2,9	*Kosten* (13), *Verlust* (10)
16	Mentale Sphäre	51	136	2,7	*Planung* (16), *Problem* (13)
17	Wahrnehmung	10	24	2,4	*Blick* (9), *Sinn* (5)
18	Seelische Sphäre	49	122	2,5	*Wahl* (10), *Bedarf* (8)
19	Sprache und Rede	96	291	3	*Frage* (22), *Rat* (19)
20	Physiologische Sphäre	35	61	1,7	*Stimme* (6), *Schmerzen* (5)
21	Verhalten und Handlungen	193	421	2,2	*Leitung* (15), *Verhältnis* (15)
22	Eigenschaften des	22	64	2,9	*Kraft* (19), *Fähigkeit* (9)

	Menschen				
23	Naturerscheinungen und Zustände	39	85	2,2	*Feuer* (11), *Sturm* (7)
24	Physikalische Eigenschaften	35	95	2,7	*Licht* (10), *Schein* (8)
25	Zeit, Alter	29	192	6,6	*Zeit* (45), *Tag* (40)
26	Kennwerte und Eigenschaften der Gegenstände	110	483	4,4	*System* (34), *Art* (25)
27	Veranstaltung, Spiel	34	113	3,3	*Spiel* (15), *Programm* (13)
28	Eigennamen	4	5	1,3	*Marathon* (2)
29	Staat, seine Attribute	32	136	4,3	*Politik* (28), *Recht* (27)
30	Dokumente, Geld	38	88	2,3	*Preis* (12), *Geld* (9)
31	Termini	62	92	1,5	*Enthalpie* (5), *Kurve* (5)
32	Sammelbezeichnungen von Menschen, Organisationen	72	288	4	*Gruppe* (22), *Partei* (16)
33	Abstrakte Begriffe	41	177	4,3	*Mittel* (26), *Möglichkeit* (12)
34	Wissenschaft, Kultur, Traditionen	47	141	3	*Punkt* (21), *Bildung* (15)

Offensichtlich werden als Hauptwörter Substantive, die Gegenstände und Instrumente, z.B. *Ulanenlanze, Zigarettenkippe, Arbeitsmaschine, Glaubensdecke, Volkswagen* (805 Belege), Personen, z.B. *Bolschewistenhäuptling, Bierkutscher, Musiklehrer, Zellenchef, Ehrenmann* (689), Raum und Ort, z.B. *Wüstenstrecke, Sandhügel, Hopfengarten, Schnittstelle, Rapsfeld* (564), Kennwerte und Eigenschaften der Gegenstände, z.B. *Wählerpotential, Religionsgleichheit, Korngröße, Bauform, Vertriebssystem* (483), Verhalten und Handlungen, z.B. *Sprachverwendung, Entwicklungshilfe, Tarifkonflikt, Liebesakt, Positivismusstreit* (421), Gebäude und Bauten, z.B. *Entlausungsbaracke, Mansardenwohnung, Atlantikwall, Schiffszimmer* (404), Tätigkeit und Aktion, z.B. *Schallübertragung, Herzfunktion, Härteübung, Verwaltungsdienst* (320), Sprache und Rede, z.B. *Ansagergeplapper, Satzaussage, Technologiegespräch, Sündenwort, Normfrage* (291), Sammelbezeichnungen von Menschen und Organisationen, z.B. *Anbieterfirma, Männerorden, Schützenverein, Parteigruppe* (288), Anzahl und Maßeinheiten, z.B. *Kleidungsfetzen, Stanniolstreifen, Gutsschicht, Satzteil, Schriftstück* (248) sowie Somatismen, z.B. *Konidienköpfchen, Drachenblut, Kalbsmaul, Gänsehaut, Dreierbrust* (234) bezeichnen, am häufigsten gebraucht.

Indem wir diese Angaben analysieren, kommen wir zum Schluss, dass semantische Unterklassen der Hauptwörter im Unterschied zu den Bestimmungswörtern den abstrakten oder konkreten Begriffen keinen Vorzug geben. In die-

sem Fall stellen wir eine fast gleiche Verteilung nach diesem Kriterium unter den größten semantischen Unterklassen fest.

Nur selten werden in dieser Funktion Eigennamen (5 Belege), Substantive der Wahrnehmung (24), possessorische (60) und physiologische (61) Sphäre bezeichnende Nomen, Eigenschaften des Menschen (64), Tiernamen (69), Bezeichnungen für Naturerscheinungen und Zustände (85), Dokumenten- und Geldbezeichnungen (88), Termini (92), physikalische Eigenschaften (95) gebraucht.

Was die Wortbildungsreihen angeht, so stellten sich als die aktivsten unter den Hauptwörtern Einheiten solcher LSU heraus:

- „Zeit, Alter", z.B. *Herzenstakt, Lichtdatum, Flugstunde, Durchlaufzeit* (Wortbildungsaktivität – 6,6 Belege), mit der längsten Wortbildungsreihe zu *Zeit*:

Feudal-, Franzosen-, Referendar-, Römer-, Zaren-, Kaiser-, Kupfer-, Bronze-, Eisen-, Stein-, Stand-, Baracken-, Uhr-, Mahl-, Aufmarsch-, Ausflugs-, Durchlauf-, Übergangs-, Bearbeitungs-, Belegungs-, Ernte-, Transport-, Trocknungs-, Dienst-, Arbeits-, Ist-, Krisen-, Lebens-, Studien-, Paarungs-, Hunger-, Verkaufs-, Kriegs-, Eis-, Still-, Fasten-, Urlaubs-, Jahres-, Soll-, Etappen-, System-, Eppler-, Sublimations-, Amts-, Schul- -zeit

- „Raum und Ort", z.B. *Asozialensiedlung, Felslandschaft, Morgenland, Bezirksstadt* (4,9), mit der längsten Wortbildungsreihe zu *Weg*:

Königs-, Esels-, Fuß-, Pappel-, Schotter-, Sand-, Stein-, Holz-, Feld-, See-, Garten-, Sonnen-, Wald-, Campus-, Heim-, Platten-, Rad-, Ziegel-, Flucht-, Strömungs-, Dienst-, Arbeits-, Entwicklungs-, Einkaufs-, Erinnerungs-, Gedanken-, Atem-, Handels-, Wirtschafts-, Karawanen-, Schul-, Kreuz- -weg

- „Kennwerte und Eigenschaften der Gegenstände", z.B. *Wählerpotential, Aufnehmerkapazität, Schuhgröße, Familienstruktur, Ausprägungsform* (4,4), mit der längsten Wortbildungsreihe zu *System*:

Experten-, Darm-, Stoff-, Sonnen-, Raster-, Waffen-, Steuerungs-, Vertriebs-, Handlungs-, Ausbildungs-, Sinn-, Wahl-, Publikations-, Kommunikations-, Medien-, Räte-, Sprach-, Belohnungs-, Betrugs-, Regelungs-, Sicherungs-, Leitungs-, Versicherungs-, Programm-, Rechts-, Wirtschafts-, Koordinaten-, Makroporen-, Stratifikations-, Absorptions-, Betriebs-, Parteien-, Hygiene-, Kultur- -system

- „Staat, seine Attribute", z.B. *Erbsteuer, Armengesetz, Volkswirtschaft, Freiheitsrecht, Sicherheitspolitik* (4,3), mit der längsten Wortbildungsreihe zu *Politik*:

Kader-, Menschen-, Welt-, Landes-, Verkehrs-, Verdrängungs-,

Regierungs-, Informations-, Medien-, Industrialisierungs-, Oppositions-, Reform-, Energie-, Tages-, Bosnien-, Europa-, Nationalitäten-, Militär-, Rechts-, Wirtschafts-, Finanz-, Steuer-, Bündnis-, Personal-, Bundes-, Familien-, Sicherheits-, Technologie- -politik

- „Gebäude und Bauten", z.B. *Geräteschuppen, Stadttheater, Nachtquartier, Bauhütte, Jagdzimmer, Landhaus* (4,1), mit der längsten Wortbildungsreihe zu *Haus*:

Grenzer-, Patrizier-, Repräsentanten-, Wirts-, Herrscher-, Pförtner-, Hexen-, Zeug-, Eltern-, Kranken-, Abgeordneten-, Schützen-, Führer-, Königs-, Nachbar-, Bauern-, Bürger-, Gast-, Bienen-, Holz-, Lager-, Land-, Court-, Giebel-, Treppen-, Büro-, Block-, Geräte-, Zucht-, Geburts-, Rat-, Johannes-, Gemeinde-, Miets-, Steuer-, Gewerkschafts-, Bank-, Pfarr-, Schul-, Volks- -haus

- „Sammelbezeichnungen von Menschen, Organisationen", z.B. *Eingeborenenfarm, Lagerkomitee, Landvolk, Randgruppe* (4), mit der längsten Wortbildungsreihe zu *Gruppe*:

Ordner-, Reporter-, Besucher-, Abgeordneten-, Christen-, Menschen-, Muskel-, Sessel-, Dreier-, Reise-, Bau-, Problem-, Planungs-, Studien-, Kommando-, Schulungs-, Kampf-, Rand-, Wert-, Tarif-, Partei-, Macht- -gruppe

- „Anzahl, Maßeinheiten", z.B. *Skatpartie, Frauenquote, Systemkomponente, Leistungsgrad, Lebensteil* (3,7), mit der längsten Wortbildungsreihe zu *Stück*:

Gesellen-, Meister-, Achsel-, Mund-, Kleidungs-, Schmuck-, Mist-, Papp-, Holz-, Grund-, Wald-, Kuchen-, Zucker-, Anschluß-, Bruch-, Beute-, Fund-, Bestimmungs-, Beweis-, Erinnerungs-, Schrift-, Probe-, Pracht-, Ausstellungs-, Kunst- -stück

- „Veranstaltung, Spiel", z.B. *Bestattungsfeierlichkeit, Rettungsaktion, Obstmarkt, Wortspiel* (3,3), mit der längsten Wortbildungsreihe zu *Spiel*:

Kinder-, Flossen-, Mienen-, Körper-, Kneipen-, Gebärden-, Wurf-, Trauer-, Wort-, Krieg-, Kräfte-, Katastrophen-, Vabanque-, Gesellschafts-, Glücks- -spiel

- „Dasein", z.B. *Kopfgeburt, Waldsterben, Ausgangsschmelze, Pesttod, Wissenschaftsgeschichte* (3,2), mit der längsten Wortbildungsreihe zu *Geschichte*:

Indianer-, Mädchen-, Weiber-, Menschen-, Löwen-, Insel-, Dorf-,

Welt-, Stadt-, Landes-, Kirchen-, Haus-, Glocken-, Entwicklungs-, Lebens-, Liebes-, Kriegs-, Natur-, Zeit-, Nibelungen-, Wirtschafts-, Dollar-, Menschheits-, Soziologie-, Wissenschafts-, Kunst- -geschichte

Unüblich sind Wortbildungsreihen für Eigennamen (1,3 Belege), Termini (1,5), Tiernamen (1,6) sowie für Bezeichnungen der physiologischen Sphäre (1,7).

Die angeführten Angaben erlauben uns unter den lexikalisch-semantischen Unterklassen von Hauptwörtern zwei Gruppen auszusondern:

1. Häufiggebrauchte LSU wegen ihrer großen Einheitsmenge, m.a.W. ihres großen Umfangs: LSU „Gegenstände und Instrumente", LSU „Person", LSU „Verhalten und Handlungen", LSU „Tätigkeit, Aktion", LSU „Sprache und Rede", LSU „Somatismen".

2. Häufiggebrauchte LSU, denen hohe Wortbildungsaktivität ihrer Einheiten eigen ist: LSU „Raum und Ort", LSU „Kennwerte und Eigenschaften der Gegenstände", LSU „Gebäude und Bauten", LSU „Sammelbezeichnungen von Menschen, Organisationen", LSU „Anzahl, Maßeinheiten".

Der Gebrauch von LSU der Substantiv-Hauptwörter sieht für verschiedene Stile (siehe Tab. 5 und Abb. 2) im Allgemeinen ähnlich aus, wovon auch die Werte der Rangkorrelation zeugen (siehe Tab. 6). Das betrifft vor allem Eigennamen, die in allen Stilen (Rang: 34) gleich wenig gebraucht werden; die Nomen der Bewegung (Rang: 16-15,5-15); Dokumente und Geld bezeichnende Substantive (Rang: 24-23,5-25,5); sehr oft kommen in allen Stilen Substantive des Raumes und des Orts (Rang: 3-5-4) sowie Kennwerte und Eigenschaften der Gegenstände (Rang: 5-6-1) und selbst Gegenstände und Instrumente (Rang: 1-4-6) vor; viel seltener aber Tiernamen (Rang: 27-30-28,5), Wahrnehmungssubstantive (Rang: 33-32-30), Begriffe aus der physiologischen Sphäre (Rang: 29-26,5-27) und physikalische Eigenschaften (Rang: 23-28-24).

Ein großer Unterschied charakterisiert solche Unterklassen, wie Termini, die in der Belletristik kaum gebraucht werden (Rang 32), aber in der Publizistik (Rang 21) schon öfter vorkommen; diese sind vor allem im wissenschaftlichen Stil (Rang 8) sehr gebräuchlich. Fast dasselbe betrifft die Substantive der mentalen Sphäre (Rang: 28-13,5-10) und abstrakte Begriffe (Rang: 20,5-18,5-5). Somatismen hingegen sind typisch für die Belletristik (Rang 6), im Unterschied zu den beiden anderen Stilen (Rang: 22-25,5); dies gilt auch für Essen und Getränke bezeichnende Substantive (Rang: 15-33-33). Die Nomen mit dem Sem „Staat, seine Attribute" kommen vor allem in den Zeitungen vor (Rang 8), viel seltener im wissenschaftlichen Stil (Rang 18) und in der Belletristik (Rang 26). Eine Differenz kennzeichnet den Gebrauch der Substantive, die Gebäude und Bauten benennen und vor allem in der Belletristik fixiert wurden (Rang: 4-10-21) und Pflanzenbezeichnungen, die im wissenschaftlichen Stil einen hohen Gebrauchsrang zeigen (Rang: 12-29-9).

Tabelle 5
Gebrauchsfrequenz und Ränge der LSU von Substantiv-Hauptwörtern im Modell N + N in verschiedenen Stilen

	LSU	Belletristik	Rang	Publizistik	Rang	wissenschaftlicher Stil	Rang
1	Person	497	2	188	1	35	14
2	Tiere	56	27	5	30	9	28,5
3	Somatismen	211	6	13	22	11	25,5
4	Attribute des Menschen	109	17	4	31	7	31,5
5	Pflanzen	146	12	6	29	61	9
6	Stoffe und Materialien	135	14	8	26,5	59	11
7	Raum und Ort	412	3	80	5	96	4
8	Gebäude und Bauten	348	4	44	10	25	21
9	Gegenstände und Instrumente	672	1	83	4	78	6
10	Essen und Getränke	122	15	2	33	6	33
11	Anzahl, Maßeinheiten	156	11	29	17	73	7
12	Bewegung	110	16	30	15,5	33	15
13	Tätigkeit, Aktion	160	10	50	9	117	3
14	Dasein	107	18	39	11,5	51	13
15	Possessorische Sphäre	24	31	10	25	28	19
16	Mentale Sphäre	51	28	31	13,5	60	10
17	Wahrnehmung	14	33	3	32	8	30
18	Seelische Sphäre	73	22	30	15,5	21	23
19	Sprache und Rede	208	7	66	7	31	17
20	Physiologische Sphäre	43	29	8	26,5	10	27
21	Verhalten und Handlungen	188	8	126	2	119	2
22	Eigenschaften des Menschen	36	30	11	23,5	22	22
23	Naturerscheinungen und Zustände	63	25	15	20	9	28,5
24	Physikalische Eigenschaften	72	23	7	28	17	24
25	Zeit, Alter	140	13	39	11,5	27	20
26	Kennwerte und Eigenschaften der Gegenstände	214	5	67	6	207	1
27	Veranstaltung, Spiel	78	19	31	13,5	7	31,5
28	Eigennamen	4	34	1	34	0	34

29	Staat, seine Attribute	61	26	56	8	29	18
30	Dokumente, Geld	66	24	11	23,5	11	25,5
31	Termini	18	32	14	21	62	8
32	Sammelbezeichnungen von Menschen, Organisationen	162	9	101	3	32	16
33	Abstrakte Begriffe	75	20,5	24	18,5	86	5
34	Wissenschaft, Kultur, Traditionen	75	20,5	24	18,5	56	12

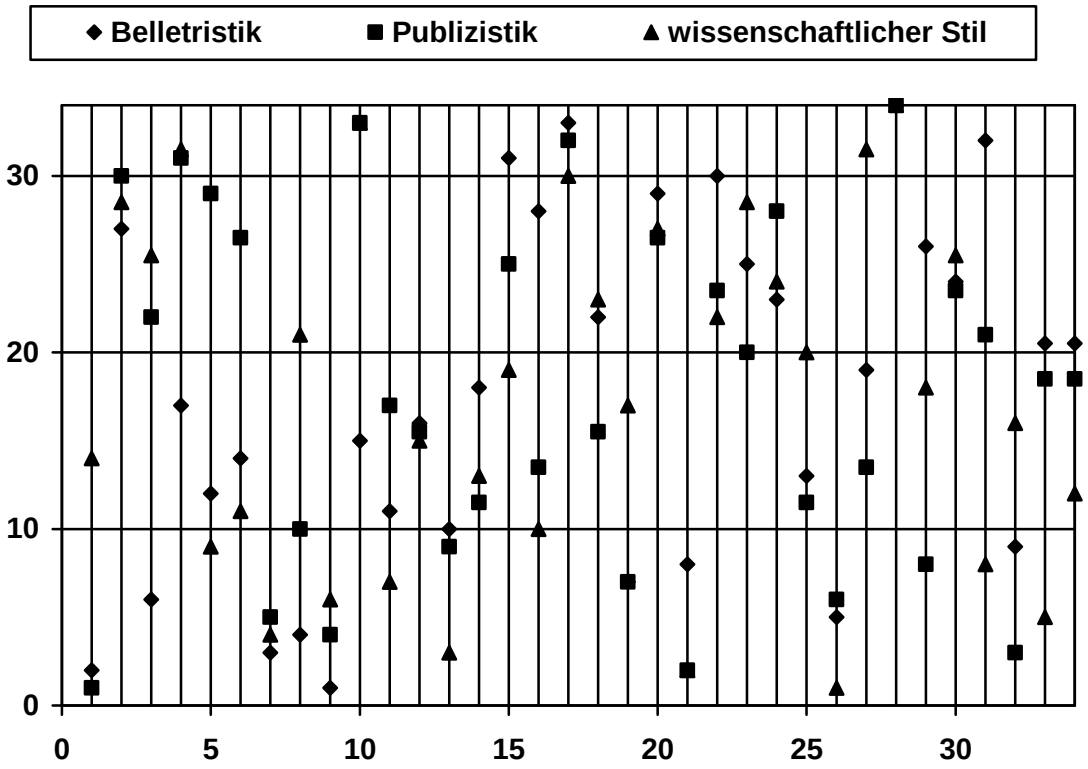


Abb. 2. Rangkorrelation der LSU von Substantiv-Hauptwörtern im Modell N + N in verschiedenen Stilen

Tabelle 6
Ähnlichkeit der Stile im Gebrauch der LSU
von Substantiv-Hauptwörtern im Modell N + N

	Belletristik	Publizistik	Wissenschaftlicher Stil
Belletristik		0,647	0,480
Publizistik			0,622
Wissenschaftlicher Stil			

Wie wir sehen, ist die Ähnlichkeit der Stile in diesem Fall viel höher als für Bestimmungswörter. Alle drei Paare erweisen effiziente Korrelationswerte bei P = 0,01. Die Tendenz zur größeren Ähnlichkeit des Paars „Belletristik-Publizistik“ ist aber auch hier zu beobachten.

5. Vergleichscharakteristiken der LSU von Substantiv-Bestimmungs- und -Hauptwörtern im Modell N + N

Die Produktivität einer lexikalischen Einheit, die in einem Kompositum mit dem Modell N + N gebraucht wird, hängt kaum von ihrer Rolle als Bestimmungs- oder Hauptwort ab (Matskulyak 2006). Um den Einfluss dieses Faktors auf den Gebrauch bestimmter lexikalisch-semantischer Unterklassen festzustellen, bringen wir alle substantivischen Bestimmungs- und Hauptwörter ihren LSU nach in Tabelle 7.

Tabelle 7
Quantitative Verteilung der LSU von Substantiv-Bestimmungswörtern und -Hauptwörtern

№	LSU	Bestimmungswort		Hauptwort	
		Zahl	Rang	Zahl	Rang
1.	Person	604	1	689	2
2.	Tiere	205	14	69	29
3.	Somatismen	333	8	234	11
4.	Attribute des Menschen	67	31	120	23
5.	Pflanzen	317	9	209	12
6.	Stoffe und Materialien	489	4	196	13
7.	Raum und Ort	521	2	564	3
8.	Gebäude und Bauten	305	10	404	6
9.	Gegenstände und Instrumente	519	3	805	1
10.	Essen und Getränke	162	17	129	21
11.	Anzahl, Maßeinheiten	119	24	248	10
12.	Bewegung	138	20	164	17
13.	Tätigkeit, Aktion	420	5	320	7
14.	Dasein	173	16	181	15
15.	Possessorische Sphäre	57	32	60	32
16.	Mentale Sphäre	107	26	136	19,5
17.	Wahrnehmung	28	34	24	33
18.	Seelische Sphäre	130	23	122	22
19.	Sprache und Rede	249	11	291	8
20.	Physiologische Sphäre	75	30	61	31
21.	Verhalten und Handlungen	373	7	421	5

22.	Eigenschaften des Menschen	39	33	64	30
23.	Naturerscheinungen und Zustände	158	18	85	28
24.	Physikalische Eigenschaften	91	29	95	25
25.	Zeit, Alter	219	12	192	14
26.	Kennwerte und Eigenschaften der Gegenstände	200	15	483	4
27.	Veranstaltung, Spiel	96	28	113	24
28.	Eigennamen	131	21,5	5	34
29.	Staat, seine Attribute	211	13	136	19,5
30.	Dokumente, Geld	102	27	88	27
31.	Termini	131	21,5	92	26
32.	Sammelbezeichnungen von Menschen, Organisationen	379	6	288	9
33.	Abstrakte Begriffe	144	19	177	16
34.	Wissenschaft, Kultur, Traditionen	114	25	141	18

Wie wir sehen, wird die allgemeine Tendenz zur Dominierung bestimmter LSU ungeachtet ihrer Funktion im Kompositum bewahrt. So gehören in beiden Fällen zu den häufigsten Unterklassen LSU „Person", LSU „Raum und Ort", LSU „Tätigkeit, Aktion", LSU „Sammelbezeichnungen von Menschen, Organisationen", LSU „Verhalten und Handlungen", LSU „Somatismen", LSU „Gebäude und Bauten" und LSU „Sprache und Rede".

Gleichzeitig fungieren hochfrequente Bestimmungswörter zur Bezeichnung der Stoffe und Materialien, Pflanzen sowie der Zeit und des Alters viel seltener als Hauptwort. Und umgekehrt treffen wir Kennwerte und Eigenschaften der Gegenstände bezeichnende Substantive sowie Anzahl und Maßeinheiten ganz oft als Hauptwörter und eher selten als Bestimmungswörter.

Seltene Bestimmungs- sowie Hauptwörter sind Einheiten, die Wahrnehmungen, Eigenschaften des Menschen und possessorische und physiologische Sphären bezeichnen. Bemerkenswert ist der rare Gebrauch der Attribute des Menschen als Bestimmungswort, wobei in der Funktion des Hauptwortes sie viel häufiger funktionieren. Ungeeignete als Hauptwort Naturerscheinungen- und Zustandsbegriffe sowie Tier- und Eigennamen kommen viel öfter als Bestimmungswörter vor.

Eine große Menge von lexikalischen Einheiten in der LSU kennzeichnet sowohl Bestimmungs-, als auch Hauptwörter mit dem Sem „Person", „Gegenstände und Instrumente", „Verhalten und Handlungen" und „Sprache und Rede". Ihrerseits ist für die LSU „Raum und Ort", „Sammelbezeichnungen von Menschen, Organisationen", „Gebäude und Bauten" in beiden Funktionen eine hohe Produktivität der Lexeme charakteristisch.

Einige Unterschiede sind für die LSU „Tätigkeit, Aktion" und „Somatismen" zu erwähnen, die als Bestimmungswörter dank der hohen Produktivität

ihrer Einheiten und als Hauptwörter dank dem großen Umfang der Lexeme hochfrequent sind.

Die nächste Graphik (siehe Abb. 3) veranschaulicht uns die Verteilung der LSU in beiden Funktionen. Auf der *x*-Achse werden die LSU-Nummer angegeben und auf der *y*-Achse die Ränge der LSU.

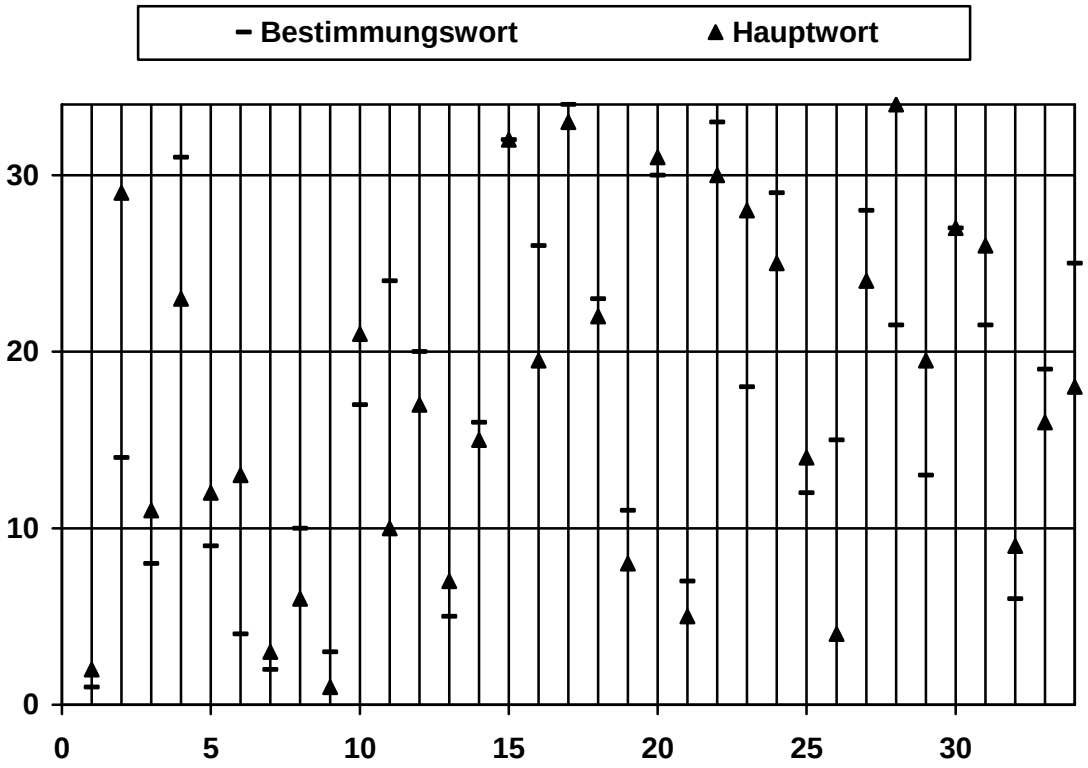


Abb. 3. Gebrauch der LSU von Substantiv-Bestimmungswörtern und –Hauptwörtern

Die Korrelationsanalyse der Daten in der Tab. 7 erlaubte uns den Ähnlichkeitsgrad zwischen diesen zwei Verteilungen zu bestimmen. Der Koeffizientenwert beträgt 0,810 ($P_{0,01}$ = 0,44), was die geahnte Ähnlichkeit bestätigt und uns behaupten lässt, dass der Gebrauch von verschiedenen Substantiv-LSU mit Ausnahme von einigen Einzelfällen wenig davon abhängt, in welcher funktionalen Rolle sie in einem Kompositum fungieren.

6. Schlussfolgerungen

Ein Fazit aus der durchgeführten Untersuchung ziehend können wir feststellen, dass das Funktionieren der LSU von Bestimmungs- und Hauptwörtern in Zusammensetzungen mit dem Modell N + N durch eine Ähnlichkeit gekennzeichnet ist.

Ein gewisser Unterschied ist zu bemerken, wenn wir die häufigsten Unterklassen beobachten: als Bestimmungswörter kommen gewöhnlich die konkrete Begriffe und viel seltener Abstraktionen, wobei die Hauptwörter eine fast gleiche Verteilung nach diesem Kriterium unter den größten semantischen Unterklassen feststellen lassen.

In beiden Fällen sprechen wir von zwei Kategorien der häufig gebrauchten semantischen Unterklassen:

1) dank der Vielfältigkeit und Fülle ihrer lexikalischen Einheiten;

2) und dank der hohen Wortbildungsaktivität ihrer Einheiten.

Die Vergleichsanalyse erlaubt uns zu behaupten, dass folgende LSU: „Person", „Gegenstände und Instrumente", „Verhalten und Handlungen", „Sprache und Rede" ausschließlich der ersten Kategorie zugeordnet sind. Dabei betrifft es sowohl Bestimmungs-, als auch Hauptwörter. Diese beiden funktionieren auch gleich, aber genießen dabei eine hohe Wortbildungsaktivität in den LSU: „Raum und Ort", „Sammelbezeichnungen von Menschen, Organisationen", „Gebäude und Bauten".

Unter diesen hochfrequenten Unterklassen gibt es auch solche, die verschiedene Kategorien in verschiedenen Funktionen bevorzugen. Als Bestimmungswörter sind die Einheiten der LSU „Tätigkeit, Aktion" und „Somatismen" hochaktiv und als Hauptwörter erweisen diese Unterklassen einen breiten lexikalischen Bestand.

Der Gebrauch von verschiedenen LSU in drei untersuchten Stilen zeigt uns positive Korrelationswerte, die eine höhere Ähnlichkeit für das Paar „Belletristik-Publizistik" ergeben. Es muss aber betont werden, dass der Grad der Korrelation für Hauptwörter stärker ist.

Wie für einzelne lexikalische Einheiten, so auch für die LSU im Allgemeinen prägt die funktionale Rolle (Bestimmungs- oder Hauptwort) ihren Gebrauch kaum. Mit Ausnahme von einigen Einzelfällen wird eine hohe Ähnlichkeit in ihrem Funktionieren bestätigt.

Literatur

Braun, P. (1998). *Tendenzen in der deutschen Gegenwartssprache: Sprachvarietäten*. Stuttgart-Berlin-Köln: Kohlhammer.

Godglück, P. (1997). Nominalkomposita in Texten – Textinformation in Nominalkomposita. Einige Aspekte ihrer Semantik. *Grazer Linguistische Studien 47, 21-40.*

Henzen, W. (1957). *Deutsche Wortbildung*. Tübingen: Max Niemeyer Verlag.

Koeder, A. (1999). *Von Ferdinand de Saussure zu einer formalen diachronischen Semantik*. Konstanz. http://deposit.ddb.de/cgi-bin/dokserv?idn=958128987& dok_var=d1&dok_ext=pdf&filename=958128987.pdf.

Langer, S. (1998). Zur Morphologie und Semantik von Nominalkomposita. *Tagungsband der 4. Konferenz zur Vorbereitung natürlicher Sprache (KONVENS): 83-97.*

Lewizkij, V., Matskulyak, Y. (2009). Semantische Kombinierbarkeit von Komponenten in der Struktur der deutschen Komposita. *Glottometrics 19, 14–41.*

Matskulyak, Y. (2006). Vžyvannja imennykiv u skladnyh konstrukcijah typu N + N. *Naukovyj visnyk Černiveckoho universytetu: Vyp. 319-320: Hermanska filologija 57–69.*

Meyer, R. (1993). *Compound Comprehension in Isolation and Context. The contribution of conceptual and discourse knowledge to the comprehension of German novel noun-noun compounds*. Tübingen: Niemeyer.

Motsch, W. (1999). *Deutsche Wortbildung in Grundzügen*. Berlin/New York: de Gruyter.

Murjassov, R.Z. (1980). *Slovoobrazovatelnaja sistema sovremennogo nemeckogo jazyka*. Ufa: Izdatel´stvo Baškirskogo universiteta.

Murphy, G.L. (1988). Comprehending complex concepts. *Cognitive Science 12, 529-562.*

Perebyjnis, V.I. (2002). *Statystyčni metody dlja lingvistiv*. Vinnycja: Nova knyha.

Springer, K., Murphy, G.L. (1992). Feature availability in conceptual combination. *Psychological Science 3, 111-117.*

Tafreschi, A. (2006). *Zur Benennung und Kategorisierung alltäglicher Gegenstände: Onomasiologie, Semasiologie und kognitive Semantik.* Kassel: Kassel University Press.

Taylor, J.R. (2003). *Linguistic Categorization*. Oxford: Oxford Univ. Press.

Ungerer, F., Schmid, H.-J. (1996). *An Introduction to Cognitive Linguistics*. London: Longman.

Vandermeeren, S. (1999). Semantische Analyse deutscher Substantiv-Komposita: Drei Untersuchungsmethoden im Vergleich. *Leuvense Bijdragen 88, 69-94.*

Vassiljev, L.M. (1990). *Sovremennaja lingvističeskaja semantika*. Moskva: Vyssšaja škola.

Kelih, E. et al. (eds.), Issues in Quantitative Linguistics, Vol. 2, 170-188

Classifying parts-of-speech systems by their quantitative properties

Relja Vulanović

1. Introduction

In Hengeveld's (1992) functional approach (see also Hengeveld et al. 2004, Hengeveld and Rijkhoff 2005, Rijkhoff 2007, Hengeveld and van Lier 2010), word classes differ between themselves according to what propositional functions (or slots) they can fulfill. Only four propositional functions are considered: P – the head of the predicate phrase, R – the head of the referential (noun) phrase, p – the modifier of the predicate phrase, and r – the modifier of the referential phrase. Verbs, for instance, can only function as P and the main function of nouns, adjectives, and manner adverbs[1] are R, r, and p, respectively. In addition to these four main word classes that we are used to, languages can have various other word classes. An example is the word class of *contentives*, found in Samoan (Hengeveld et al. 2004). This word class has all four propositional functions.

Languages use different combinations of word classes to fulfill their propositional functions. The number of propositional functions is in some cases less than 4. For instance, Tagalog is missing the slot for the modifier of the predicate phrase (ibid.). Different combinations of word classes define different parts-of-speech (PoS) systems. Seven main PoS system types are proposed in Hengeveld (2002). This classification is extended in Hengeveld et al. (2004) to include six additional intermediate types. However, it is shown later on (Hengeveld and van Lier 2010) that there exist natural languages with PoS systems which do not fit within the original classification. For examples of different PoS systems in natural languages, the reader is referred to the above-mentioned papers.

Since linguists define word classes in many different ways (see, for instance, Givón 1993, Vogel and Comrie 2000, Baker 2003, Rijkhoff 2007), Hengeveld's definition of PoS systems and their classification are not universally accepted (Ansaldo et al. 2010). Hengeveld's approach to PoS systems is nevertheless interest-

[1] Adverbs other than manner adverbs are not considered because they usually modify the whole sentence and not just the head of the predicate phrase.

ing and has enough merit to be discussed further. It is appealing to quantitative and mathematical linguists because it can be mathematically generalized and analyzed.

The 7 main PoS system types of Hengeveld's classification are simply labeled as types 1-7. Types 1-4 are related to some quantitative characteristics of PoS systems. More specifically, the type number is the number of word classes used, but only when all 4 propositional functions exist in the system. Therefore, this part of the classification scheme is quantitatively relevant but not quite complete. Tagalog and Samoan, for instance, are both classified as type 1 languages although they do not have the same number of propositional functions (see above). Moreover, the labels of types 5-7 have nothing to do with any quantitative property of the PoS system. Because of all this, a two-dimensional classification of PoS systems is proposed in Vulanović (2008). In this classification scheme, each PoS system is labeled by *k.n*, where *k* is the number of word classes and *n* is the number of propositional functions. The viability of the two-dimensional classification system has been confirmed by different correlation analyses (Vulanović and Köhler 2009, Vulanović 2010, Vulanović and Miller 2010).

However, the two-dimensional classification, although quantitatively sound, is not fine enough to distinguish between some PoS systems. With the new PoS system types reported in Hengeveld and van Lier (2010), it is possible to have structurally different PoS systems with the same numbers *k* and *n*. In other words, some systems of type *k.n* have subtypes, the difference between which is not only formal, but essential. Such subtypes have already been considered in Vulanović (2008). Their classification labels consist of the numerical part *k.n*, followed by one or two letters referring to the propositional functions, the way they are fulfilled in the system; see Vulanović (2010) and Vulanović and Miller (2010) as well. This means that any quantitative investigation of PoS systems, which involves *k* and *n*, has to lump together all subtypes of the same type. Vulanović and Miller (2010) is a particularly illustrative example of this.

The purpose of the present paper is to investigate the possibilities for obtaining fully quantitative classification labels for all the PoS system types and subtypes of interest. It is shown that this is possible to do and four different classification schemes are proposed. Two of those use three-number labels, whereas the other two use four-number labels. All four schemes preserve *n* as part of the label, but the word classes are divided into two or three subclasses and the corresponding two or three numbers are used instead of *k*. As a way of confirming the adequacy of the new classification schemes, it is shown that there exists a correlation between the grammar efficiency of PoS systems and their new labels. A motivation for this comes from Vulanović and Miller (2010), where a correlation of this kind is established for the two-number classification. Vulanović and Miller (ibid.) use a three-dimensional generalization of the sigmoid to represent this nonlinear

correlation and they provide an explanation why this is an appropriate model. In the same way, four- and five-dimensional generalizations of the sigmoid are used here.

Two kinds of grammar-efficiency values are considered – those corresponding to the free word order and those corresponding to the fixed word order. Grammar efficiency is calculated according to the formula from Vulanović (2003, 2007). This formula has been used in Vulanović (2008, 2009) and Vulanović and Miller (2010) in investigations of PoS systems.

The paper is organized as follows. All theoretically possible word classes within Hengeveld's framework are presented in Section 2. PoS systems are then discussed in Section 3. After some appropriate assumptions are applied, 17 PoS system types and subtypes are left for consideration, 13 of them being structurally different. Each of those receives a unique label in the three- and four-number classification schemes. In Section 4, the concept of grammar efficiency is briefly explained and the results of calculating grammar-efficiency values for each PoS system are presented. These results are used in Section 5 as the data to which the multidimensional generalizations of the sigmoid are fitted. Finally, concluding remarks are presented in Section 6.

2. Word classes

As already mentioned, languages may have less than 4 propositional slots. The number of propositional slots, *n*, can be 1, 2, 3, or 4. Table 1 shows all possible combinations of propositional slots (Vulanović 2008). The rows for $n = 4, 3, 2$ can be explained by saying that the heads P and R are the obligatory slots, whereas the modifier slots p and r are optional. The row for $n = 1$ is added for completeness in order to include the unattested type 1 PoS system of the Hengeveld (1992) classification.

Table 1
Possible combinations of propositional slots

n	P	R	r	p
4	+	+	+	+
3	+	+	+	-
3	+	+	-	+
2	+	+	-	-
1	+	-	-	-

Keeping the combinations of Table 1 in mind, we arrive at the inventory of all possible word classes, presented in Tables 2-4. Each of the cases, $n = 4$, $n = 3$, or $n = 2$, is represented in a separate table. In the theoretical case $n = 1$, verbs form the only word class. The tables also introduce the notation which is used for different word classes. The label X in Table 2 is used generically – it does not stand for the same word class in each row in which it occurs; note also that these word classes are left unnamed. Each word class is defined by the propositional slots it can fill.

Verbs, nouns, adjectives, and manner adverbs are the only *rigid* word classes and all other word classes are *flexible*. A rigid word class is specialized for one and only one propositional slot, whereas a flexible word class can fill two or more different propositional slots.

Table 2
Word classes in the presence of 4 propositional slots

Word class	P	R	r	p
Verbs	V	-	-	-
Nouns	-	N	-	-
Adjectives	-	-	a	-
Manner adverbs	-	-	-	m
Heads	H	H	-	-
Predicatives	$\mathcal{P}$	-	-	$\mathcal{P}$
Nominals	-	$\mathcal{N}$	$\mathcal{N}$	-
Modifiers	-	-	M	M
-	X	-	X	-
-	-	X	-	X
Non-verbs	-	Λ	Λ	Λ
Non-nouns	Z	-	Z	Z
-	X	X	X	-
-	X	X	-	X
Contentives	C	C	C	C

Table 3
Word classes in the presence of 3 propositional slots

Word class	P	R	r **or** p
Verbs	V	-	-
Nouns	-	N	-
Adjectives or manner adverbs	-	-	a m
Heads	H	H	-
Non-verbs	-	Λ	Λ
Non-nouns	Z	-	Z
Contentives	C	C	C

Table 4
Word classes in the presence of 2 propositional slots

Word class	P	R
Verbs	V	-
Nouns	-	N
Contentives (heads)	C	C

3. PoS systems

The above word classes are combined to create different PoS systems. Only those systems are considered here that satisfy the following three assumptions:

Assumption 1. The possible combinations of propositional slots in PoS systems are those presented in Table 1.
Assumption 2. Each existing propositional slot is filled by one and only one word class.
Assumption 3. The only word classes a PoS system can use in both predicate and referential slots are C, Λ, and M, as defined in Tables 2-4.

Assumption 3 is constraint C2 of Vulanović (2010). It prohibits word classes H, X, and Z from being used in PoS systems. The assumptions leave the 17 PoS systems given in Table 5. All PoS systems that are eliminated by Assumptions 1-3 are unattested (ibid.); see also Hengeveld and van Lier (2010). Some of the remaining ones are also unattested and marked by an asterisk. Even the attested PoS

system types in Table 5 represent ideal structures that natural languages only approximate (Rijkhoff 2007: 718).

Table 5
PoS systems satisfying Assumptions 1-3

PoS system type	**Subtype**		**Hengeveld's type**	P	R	r	p
1.4	-		1	C	C	C	C
2.4	P		2	V	Λ	Λ	Λ
	Pp = Rr		-	𝒫	𝔑	𝔑	𝒫
3.4	HM	*Pp	-	𝒫	N	a	𝒫
		Rr	-	V	𝔑	𝔑	m
	rp		3	V	N	M	M
4.4	-		4	V	N	a	m
1.3	r		-	C	C	C	-
	*p		-	C	C	-	C
*2.3	*HM	*Pp	-	𝒫	N	-	𝒫
		*Rr	-	V	Λ	Λ	-
	*Rp		-	V	Λ	-	Λ
3.3	r		5	V	N	a	-
	*p		-	V	N	-	m
1.2	-		-	C	C	-	-
2.2	-		6	V	N	-	-
*1.1	-		7	V	-	-	-

Each PoS system type is labeled as *k.n*, where *k* is the number of word classes. Some of these types have subtypes which are additionally marked by one or two propositional-slot labels. Two slot labels indicate that there is one word class filling exactly the two indicated propositional slots. A single head label means that the corresponding propositional slot is filled by a rigid word class and the remaining slots share another word class. A single modifier label indicates the only existing modifier slot. Subtypes 3.4Pp and 3.4Rr have equivalent structures (one word class for either both predicate or both referential slots and rigid word classes for the other two slots) and this is why a joint label 3.4HM is used to represent them both. Similarly, 2.3Pp and 2.3Rr are equivalent and are labeled jointly as 2.3HM. The HM part of these labels is meant to indicate that the head and the modifier belong to the same phrase (Vulanović 2008). Other pairs of equivalent subtypes are 1.3r and 1.3p,

as well as 3.3r and 3.3p – in this case 1.3 and 3.3, respectively, suffice as the corresponding joint labels.

There are 5 rigid PoS systems in Table 5; all others are flexible. A PoS system is rigid if all its word classes are rigid. In rigid PoS systems, $k = n$. On the other hand, a flexible PoS system has at least one word class which is flexible, thus $k < n$. Of the 17 PoS systems in Table 5, 7 coincide with the original Hengeveld (1992) PoS system types, as indicated in the table. This original classification is extended in Hengeveld et al. (2004) to intermediate PoS system types, but these are not considered here. In particular, flexible intermediate types do not satisfy Assumption 2.

In addition to Assumptions 1-3, another constraint on PoS systems is considered in Vulanović (2010). This constraint, labeled there as C3, requires that if a language has nouns, it must have verbs (cf. Hengeveld and van Lier 2010). Although it correctly eliminates the unattested types 3.4Pp and 2.3Pp, C3 is not as powerful as Assumption 3. Overall, Assumption 3 correctly eliminates 11 unattested PoS system types and C3 only 5, 3 of which are covered by Assumption 3. Moreover, even if 3.4Pp and 2.3Pp were excluded, their equivalent types 3.4Rr and 2.3Rr, respectively, would remain under consideration and the results of the present analysis would not change. This is why constraint C3 is not taken into account here.

Table 5 shows that there are 10 main PoS system types but 13 non-equivalent structures. This is because type 2.4 has two essentially different subtypes, 2.4P and 2.4Pp; type 3.4 – 3.4HM and 3.4rp; and type 2.3 – 2.3HM and 2.3Rp. Clearly, the labels used for these subtypes are not entirely quantitative – the counts k and n apply jointly to both subtypes within the same type. This is why in all quantitative investigations, which involve k and n, the subtypes of the same type have to be combined together. For instance, this is done in Vulanović and Miller (2010), where the correlation between the PoS system type label $k.n$ and grammar efficiency is analyzed. It is therefore of interest to find some quantitative characteristics that will make a distinction between the subtypes of the same type.

Four possibilities are considered here. They can be viewed as new numerical labels for the 13 PoS systems, since each system is represented in a unique way. Table 6 shows these labels, denoted by L1-L4. The format of each label in Table 6 is either k_1k_2n or $k_1k_2k_3n$, which means that the last digit, like in the two-number labeling, indicates the number of propositional slots. The leading digits are such that $k_1 + k_2 = k$, i.e., $k_1 + k_2 + k_3 = k$. This is because word classes are divided into two or three disjoint groups and the count within each group is included in the label. In label L1, k_2 is the number of word classes that are used both in a predicate slot (P or p) and in a referential slot (R or r). Then, k_1 is the number of the remaining word classes ($k_1 = k - k_2$), which comprise the rigid classes and those flexible classes which are only used within either predicate or referential phrase, but not in both phrases. L2 is a refinement of L1: the count k_1 of label L1 is split between the rigid

word classes (first digit of label L2) and the flexible classes which do not cross the boundary between the predicate phrase and the referential phrase (second digit of label L2). In label L3, k_2 is the number of word classes with both head and modifier functions within the same phrase (either predicate or referential). The relationship between L4 and L3 is analogous to that between L2 and L1. The first digit of label L4 is the number of rigid word classes, the second digit is the number of flexible word classes that are not used as both the head and modifier within the same phrase, and the third digit is the same as k_2 in L3.

Table 6
Four numerical labels for 13 PoS systems

PoS system	L1	L2	L3	L4
1.4	014	0014	014	0014
2.4P	114	1014	114	1014
2.4Pp	204	0204	024	0024
3.4HM	304	2104	214	2014
3.4rp	214	2014	304	2104
4.4	404	4004	404	4004
1.3	013	0013	013	0013
2.3HM	203	1103	113	1013
2.3Rp	113	1013	203	1103
3.3	303	3003	303	3003
1.2	012	0012	102	0102
2.2	202	2002	202	2002
1.1	101	1001	101	1001

It is established in Vulanović and Miller (2010) that there exists a correlation between *k.n*, as the PoS system type, and the efficiency of the corresponding grammatical structure. The correlation is modeled by a three-dimensional generalization of the sigmoid. A natural question is whether a correlation of the same kind exists when the two-number label is replaced with the new three- or four-number labels of Table 6. The rest of the paper is devoted to finding an answer to this question.

4. Grammar efficiency

The PoS system (i.e. the mapping that shows what word classes have what propositional functions), together with the permitted order of propositional functions,

forms a *grammar*. The grammatical structures considered here idealize what can be observed in natural languages, not only because PoS system types themselves represent an idealization, but also because grammatical markers are ignored. This paper is theoretical in the same sense as Vulanović (2008) or Vulanović and Miller (2010) since no linguistic sample is used in the analysis. However, it should be mentioned that grammar efficiency can be calculated in the presence of grammatical markers, the way this is done in Vulanović (2009), where a 50-language sample from Hengeveld et al. (2004) is considered.

From now on, the order of propositional functions will be simply referred to as *word order*. PoS systems can use different word orders since this is not their defining characteristics. Only continuous predicate and referential phrases are considered for simplicity. This means that the following 18 strings represent all possible word orders:

(1a) PR, RP,

(1b) PpR, pPR, RPp, RpP,

(1c) PRr, PrR, RrP, rRP,

(1d) PpRr, PpRr, pPRr, pPrR, RrPp, RrpP, rRPp, rRpP,

with an additional special case – PoS systems of type 1.1 have P as the only string. A PoS system can permit all the 18 orders in (1) or it can have a fixed word order. If $n = 4$, the minimum possible number of orders, denoted by ρ_*, is $\rho_* = 4$: one of each in (1a), (1b), (1c), and (1d). If $n = 3$, then $\rho_* = 2$: one of the two word orders in (1a) and either one in (1b), or one in (1c). Finally, $\rho_* = 1$ if $n = 2$ or $n = 1$.

The grammar efficiency, *Eff*, is calculated by the formula from Vulanović (2003, 2007), see also Vulanović (2008, 2009):

(2) $$Eff = \gamma\, Q\, \frac{n}{k}.$$

In this formula, γ is a scaling coefficient ensuring that $Eff = 1$ for maximally efficient grammars. According to the already introduced notation, n is the number of propositional slots and k is the number of word classes. The quantity Q is called the *parsing ratio* since its definition involves 3 parsing-related quantities,

(3) $$Q = (\rho - \rho_0)/\rho^*.$$

Here, ρ is the number of successful parses of all existing sentences and ρ_0 is the number of ambiguous parses. Both ρ and ρ_0 depend on the word-order rules considered. On the other hand, ρ* only depends on the PoS system type or subtype; ρ* denotes the number of parsing attempts, of which at least one is successful, applied to all permutations of all possible sentences. By a *sentence*, we mean a formal string of word classes, like VMN in a PoS system of type 3.4rp for instance. The propositional function of each word in the sentence is determined during the parsing process. Parsing[2] is considered successful if it results in at least one of the 18 strings in (1). Thus, VMN can be parsed as PpR or PrR, which means that this sentence is ambiguous unless word order is used as a disambiguation device. For instance, if the propositional functions are ordered like in pPrR, then the only interpretation of VMN is PrR. Word order is not restricted when calculating ρ* and all permutations of each sentence are considered. Moreover, attempted unsuccessful pares are counted as well. An example of this would be r- (where the minus indicates that the parse has to be abandoned), as an attempted parse of the sentence MVN (for which the only successful parse is pPR).

In this way, *Q* measures how free word order is. If ρ = ρ*, word order is completely free and then, if no sentence is ambiguous, *Q* = 1. Otherwise, *Q* < 1 either because there are some word-order rules making ρ < ρ*, or because some sentences are ambiguous. For the given PoS system, *Q* has different values for different word orders. This is why an interval [Q_*, *Q**] of *Q* values can be assigned to each PoS system. To illustrate how values of Q_* and *Q** are calculated, let us consider PoS systems types 3.4rp and 3.4Rr. Other examples can be found in Vulanović (2003, 2007, 2008).

All the possible sentences in 3.4rp are the permutations of the strings VN, VMN, and VMNM, keeping V and M together and N and M together. There are 16 sentences and only 2 are ambiguous:

VN → PR	NV → RP
VMN → PpR, PrR	VNM → PRr
NMV → RrP, RpP	NVM → RPp
MVN → pPR, r-	MNV → rRP, p-
VMNM → PpRp, PrR-	VMMN → PprR, Pr-

[2] It is assumed that parsing is done from left to right, one word at a time.

MVMN → pPrR, r- MVNM → pPRr, r-

NMVM → RrPp, RpP- NMMV → RrpP, Rp-

MNMV → rRpP, p- MNVM → rRPp, p-

There are 28 parsing attempts in all and therefore $\rho^* = 28$. The ambiguity of sentences VMN and NMV can be resolved by imposing word-order rules to eliminate one of the two possible parses in each ambiguous sentence. This is why the greatest number of unambiguous pareses, denoted by ρ', is $\rho' = 16$. If ρ is increased to 17, then ρ_0 becomes 2 and $\rho - \rho_0 < 16$. This illustrates that the inequality $\rho - \rho_0 \leq \rho'$ always holds true. Therefore, from (3) we get that $Q \leq \rho'/\rho^*$ and we set

$$Q^* = \rho'/\rho^*.$$

For the 3.4rp PoS system, $Q^* = 16/28 = 4/7$. We can see that the value of Q^* results when word order is maximally free without creating ambiguous sentences. As opposed to this, fixed word order is considered for Q_*. We have seen above that $\rho \geq \rho_*$, so we take

$$Q_* = (\rho_* - \rho_0)/\rho^*,$$

where ρ_0 is the number of ambiguous parses for the fixed word order under consideration. In 3.4rp, $\rho_* = 4$, but ambiguity cannot be avoided if the word order is fixed as PprR (or RrpP), since the sentence VMN (respectively, NMV) is then ambiguous. In this case, $\rho_* - \rho_0 = 4 - 2 = 2$, which gives $Q_* = 2/28 = 1/14$. Therefore, the interval [1/14, 4/7] is assigned to the 3.4rp PoS system. This interval contains the values of the parsing ratio Q for all word orders without ambiguous sentences. This is in general how Q_* and Q^* are calculated – ambiguous sentences are completely avoided if possible. If this is not possible, like in some fixed word orders in 3.4rp, ambiguity is then reduced to a minimum.

A different interval can be assigned to 3.4Rr, where all the possible sentences are the permutations of elements in the strings V𝔑, V𝔑𝔑, Vm𝔑, and Vm𝔑𝔑, this time keeping 𝔑 and 𝔑 together and V and m together. This gives 12 sentences which are parsed as follows:

V𝔑 → PR, Pr- 𝔑V → RP, r-

Vm𝔑 → PpR, Ppr- 𝔑Vm → RPp, r-

mV$\mathfrak{N}$ → pPR, pPr-	$\mathfrak{N}$mV → RpP, r-
V$\mathfrak{N}\mathfrak{N}$ → PRr, PrR	$\mathfrak{N}\mathfrak{N}$V → RrP, rRP
Vm$\mathfrak{N}\mathfrak{N}$ → PpRr, PprR	mV$\mathfrak{N}\mathfrak{N}$ → pPRr, pPrR
$\mathfrak{N}\mathfrak{N}$Vm → RrPp, rRPp	$\mathfrak{N}\mathfrak{N}$mV → RrpP, rRpP

From this parsing process, ρ* = 24 and ρ' = 12, so that $Q^* = 12/24 = 1/2$. Also, $\rho_* = 4$ and, no matter what fixed word order is chosen, no sentence is ambiguous. Therefore, $Q_* = 4/24 = 1/6$ and the interval [1/6, 1/2] is assigned to the 3.4Rr PoS system.

The two PoS systems considered above are subtypes of the same 3.4 type. In the efficiency formula (2), Q is the only quantity they differ by. This is because γ has the same value for all PoS systems with the same k and n (see below). Therefore, for the same word-order rules, 3.4rp and 3.4Rr have different efficiency. This is indeed supposed to be the case since parsing reveals that these two PoS system types have different structures. This is also why, in quantitative investigations, they should be treated as separate PoS systems. Different numerical labels enable this. On the other hand, the four pairs of subtypes, which are mentioned above as structurally equivalent (3.4Pp and 3.4Rr, 1.3r and 1.3p, 2.3Pp and 2.3Rr, and 3.3r and 3.3p), have equal efficiency values provided they use the same word order.

The value of γ depends on k and n, i.e. on the PoS system type. Within the family of grammars with the given k and n, we have to find the one with the greatest parsing ratio and no ambiguous sentences. Suppose such a grammar exists and let its parsing ratio be $Q^\wedge$. The efficiency of this grammar is set equal to 1, which then gives γ as $\gamma = k/(nQ^\wedge)$. Only Assumptions 1 and 2 are taken into account when looking for the maximally efficient grammar. Therefore, the class of grammars considered for finding the maximally efficient one also contains subtypes other than those presented in Table 5. For instance, the maximally efficient grammar within the 3.4 type belongs to an unattested PoS system in which

H → P, R, a → r, m → p

(Vulanović 2008). It should be pointed out that γ is not needed in Vulanović (2008). The values of Q suffice there because the comparison of efficiency values is only done between the subtypes of the same PoS system type. When, like here, all types are considered, the scaling nature of γ provides the same yardstick for measuring grammar efficiency of all PoS system types. In this sense, the grammar-efficiency formula (2) is a relative measure (Vulanović 2007).

If $k = n$, word order can be completely free without creating ambiguity, thus, $Q^{\wedge} = 1$ and $\gamma = 1$. Otherwise, if $k < n$ and the maximally efficient grammar can be found, then $\gamma > 1$, see Table 7. There is no maximally efficient grammar for the 1.4 type since ambiguity cannot be avoided in this case. For this type, γ is set equal to 1, which is otherwise the smallest possible value.

Since parsing is a highly algorithmic process, a computer program was created for calculating Q for all PoS system types satisfying Assumptions 1 and 2. This program was used to find the values of γ. The results are given in Table 7.

Table 7
The values of the coefficient γ

PoS system type	γ
1.4	1
2.4	11/8
3.4	6/5
4.4	1
1.3	5/3
2.3	7/6
3.3	1
1.2	1
2.2	1
1.1	1

With γ defined as in Table 7, we use Q_* and Q^* to calculate the following two efficiency values:

$$Eff_* = \gamma\, Q_* \frac{n}{k}, \qquad Eff^* = \gamma\, Q^* \frac{n}{k}.$$

They are presented in Table 8. For any given PoS system, Eff^* is the largest efficiency value which corresponds to the maximally free word order, whereas Eff_* corresponds to fixed word orders. Both Eff^* and Eff_* are found for grammars without ambiguous sentences if such grammars exist. If ambiguity cannot be avoided, it is reduced to a minimum when calculating Eff^* and Eff_*. Ambiguity is only present in type 1.4 (both in Eff_* and Eff^*) and in Eff_* for 2.4P and 3.4rp (the latter is discussed above). Thus, the interval $[Eff_*, Eff^*]$ is an interval of efficiency values for gram-

mars without ambiguous sentences or grammars with a minimum, unavoidable amount of ambiguity.

Each pair of equivalent systems is presented in Table 8 by one interval of efficiency values. It can be seen that some efficiency intervals reduce to a single value. This is the case if and only if $k = 1$, when word order has to be maximally restricted since it is the only disambiguation device in the absence of more word classes. This is also why PoS system types 1.3, 1.2, and 1.1 are maximally efficient and only have one efficiency value, equal to 1. If a PoS system is not maximally efficient, then there is a unique efficiency interval assigned to it.

Table 8
Efficiency values for PoS system types and subtypes

PoS system type/subtype	Eff_*	Eff^*
1.4	0.250	0.250
2.4P	0.162	0.726
2.4Pp	0.393	0.786
3.4HM	0.267	0.800
3.4rp	0.114	0.914
4.4	0.222	1.000
1.3	1.000	1.000
2.3HM	0.438	0.875
2.3Rp	0.350	0.875
3.3	0.333	1.000
1.2	1.000	1.000
2.2	0.500	1.000
1.1	1.000	1.000

5. Fitting the data

In Vulanović and Miller (2010), the 13 PoS systems in Table 8 are reduced to the 10 main types without subtypes. The data for the subtypes 2.4P and 2.4Pp are combined by making the union of the efficiency intervals and getting the interval [0.162, 0.786] to represent the 2.4 type. The same is done with subtypes 3.4HM and 3.4rp, as well as with 2.3HM and 2.3Rp. This procedure is not needed here since each PoS system in Table 8 is represented by its unique label in each of the four new classification schemes of Table 6.

Most languages tend to have either a completely free word order or a variation of a fixed word order (Comrie 1989: 88). This is why the endpoints of the efficiency intervals are only considered[3] here and their correlation to the three- and four-number labels is analyzed. When the labeling systems L1 or L3 are used, the data consist of ordered quadruples (k_1, k_2, n, y), where y is either Eff_* or Eff^*. For L2 and L4, on the other hand, the data are ordered quintuples (k_1, k_2, k_3, n, y).

All efficiency values in Table 8 can be viewed as being between two plateaus, the plateau of optimal values equal to 1 and the theoretically possible plateau of values equal to 0. No efficiency value in Table 8 equals 0 because only such grammars are considered which have a minimum amount of ambiguity, if at all. However, $Eff = 0$ if all sentences are ambiguous, which happens if $k < n$ and word order is made completely free. The two plateaus motivate the use of the multidimensional generalization of the sigmoid as a model to fit the data. This model is given by the equation

$$(4) \quad y = 1/[1 + \exp(\sum_{i=1}^{m} a_i k_i + bn + c)],$$

where $m = 2$ for labels L1 and L3, and $m = 3$ for L2 and L4.

Equation (4) is a generalization of the well-known two-dimensional sigmoid which is a curve used to model various linguistic phenomena, see Altmann (1983), Leopold (2005), Vulanović and Baayen (2007), and other references in these papers. In two dimensions, the sigmoid usually shows how some linguistic quantity changes over time. This is known as the Piotrowski or Piotrowski-Altmann Law. A three-dimensional generalization of the sigmoid is introduced in Vulanović and Köhler (2009) to model how the proportion of languages with fixed word order or grammatical markers varies in dependence on k and n. The language sample from Hengeveld et al. (2004) is used for this. The three-dimensional sigmoidal surface is a special case of equation (4), from which it follows by setting $m = 1$ and $k_1 = k$. A result similar to that of Vulanović and Köhler (2009) is obtained in Vulanović (2010), only values of k and n are modified in order to distinguish between main and intermediate PoS system types. The three-dimensional sigmoid is also used in Vulanović and Miller (2010) in the same kind of analysis as the present one, but without a separate representation of the structurally different subtypes of the same type.

[3] Recall that the Eff^* values are obtained when word order is as free as possible without creating ambiguous sentences (if this can be achieved). The Eff_* values, on the other hand, follow from a fixed word order.

The results of fitting equation (4) to the data are presented in Table 9. The values of the adjusted coefficient of multiple determination Ra^2 show that the fit is better for *Eff** then for Eff_*. With Ra^2 values over 86%, the fit for *Eff** is very good in all four labeling systems. As for Eff_*, the fit is still acceptable, with labels L1 and L2 being somewhat worse than L3 and L4. It seems that the system L3 provides the best overall fit. The results in Vulanović and Miller (2010), for the correlation when the two-number classification scheme is used, are somewhat better: $Ra^2 = 0.740$ for Eff_* and $Ra^2 = 0.951$ for *Eff**.

Table 9
Results of the fit

PoS labeling system	**Efficiency value**	a_1	a_2	a_3	b	c	Ra^2
L1	Eff_*	0.936	1.127	-	0.943	-4.703	0.680
	*Eff**	-1.630	-1.351	-	3.958	-13.619	0.861
L2	Eff_*	0.829	-0.052	-0.032	1.368	-4.914	0.685
	*Eff**	-1.950	-3.196	-4.218	3.806	-9.980	0.888
L3	Eff_*	0.710	-0.030	-	1.431	-5.212	0.726
	*Eff**	-1.686	-2.143	-	3.725	-11.838	0.878
L4	Eff_*	0.741	0.239	-0.265	1.504	-5.170	0.702
	*Eff**	-1.596	-2.060	-2.163	3.801	-12.137	0.866

6. Conclusions

This paper is concerned with some quantitative characteristics of parts-of-speech (PoS) systems. PoS systems and the word classes used in them are described following the approach from Hengeveld (1992), Hengeveld et al. (2004), and Hengeveld and van Lier (2010). After some appropriate assumptions are applied, 17 PoS systems are left for discussion. This number is reduced to 13 types or subtypes after unifying the pairs of PoS systems that have equivalent grammatical structures. In the previously introduced two-dimensional classification of PoS system types (Vulanović 2008, 2010), each type is labeled by *k.n*, where *k* is the number of word classes and *n* is the number of propositional slots (functions). In this, $1 \leq k \leq n$ and $n = 1, 2, 3, 4$. Only 10 PoS system types can be distinguished when this kind of labeling is used: 1.4, 2.4, 3.4, 4.4, 1.3, 2.3, 3.3, 1.2, 2.2, and 1.1. This is more than the 7 main types in the Hengeveld et al. (2004) classification. However, 3 of the 10 PoS system types have 2 structurally different subtypes each. Those subtypes do not

have a separate numerical representation. Thus, in Vulanović and Miller (2010), the correlation between the 10 main PoS system types and their grammar efficiency is established without the possibility to treat the subtypes of the same PoS system type separately. This paper introduces some new quantitative characteristics of PoS systems, which enable a unique numerical representation of each of the 13 PoS systems. The new quantitative characteristics are a refinement of how the word classes are counted. For instance, the rigid word classes can be counted separately from the flexible ones. It can also be counted how many word classes have both the predicate and referential functions, and how many can be used as both the head and the modifier of the same phrase (predicate or referential). Four different ways of counting subclasses of word classes are proposed. This gives four new classification schemes for the 13 PoS systems, two with three- and two with four-number labels.

In order to confirm the validity of the new labeling systems, the same kind of correlation analysis as in Vulanović and Miller (2010) is carried out. It is shown that there indeed is a correlation, modeled by equation (4), between the grammar-efficiency values considered and the three- and four-number labeling systems. Grammar efficiency is calculated using the approach from Vulanović (2003, 2007). The formula (2) assigns an interval of grammar-efficiency values to each PoS system. The values in each interval vary depending on the word-order rules. Each of the 13 PoS systems has a unique grammar-efficiency interval. Only the endpoints of the intervals are used in the correlation analysis. The left endpoints represent PoS systems with fixed word order, whereas the right endpoints correspond to systems with maximally free word order without ambiguous sentences.

The modeling equation (4) is a multidimensional generalization of the sigmoid and is justified by two plateaus of efficiency values, one with the values equal to 0 and another with the values equal to 1. A three-dimensional generalization of the sigmoid curve is introduced in Vulanović and Köhler (2009). It is used there, as well as in Vulanović (2010) and Vulanović and Miller (2010), in various analyses of PoS systems. The present model is a further generalization of the sigmoid to four and five dimensions.

It should be mentioned that another confirmation of the viability of the new multidimensional labeling systems is the fact that they can be extended to the intermediate PoS system types, as defined in Hengeveld et al. (2004). This can be done in the same way as in the case of the two-dimnesional labeling system (Vulanović 2010).

References

Altmann, G. (1983). Das Piotrowski-Gesetz und seine Verallgemeinerungen. In: Best, K.-H., Kohlhase, J. (Eds.), *Exakte Sprachwandelforschung: 54-90.* Göttingen: Herodot.

Ansaldo, U., Don, J., Pfau, R. (Eds.) (2010). *Parts of Speech: Empirical and Theoretical Advances.* Amsterdam/Philadelphia: John Benjamins.

Baker, M.C. (2003). *Lexical Categories*. Cambridge: Cambridge University Press.

Comrie, B. (1989). *Language Universals and Linguistic Typology: Syntax and Morphology*. Oxford: Blackwell.

Givón, T. (1993). *English Grammar: A Function-Based Introduction, Volume 1*. Philadelphia: John Benjamins.

Hengeveld, K. (1992). Parts of speech. In: Fortescue, M., Harder, P., and Kristoffersen, L. (Eds.), *Layered Structure and Reference in Functional Perspective*: *29-55.* Amsterdam/Philadelphia: John Benjamins.

Hengeveld, K., Rijkhoff, J. (2005). Mundari as a flexible language. *Linguistic Typology 9, 406–431*.

Hengeveld, K., Rijkhoff, J., Siewierska, A. (2004). Parts-of-speech systems and word order. *Journal of Linguistics 40, 527–570.*

Hengeveld, K., Lier, E.v. (2010). An implicational map of parts of speech. *Linguistic Discovery 8, 129-156.*

Leopold, E. (2005). Das Piotrowski-Gesetz. In: Köhler, R., Altmann, G., and Piotrowski, R. (Eds.), *Quantitative Linguistics: An International Handbook*: *627-633.* Berlin/New York: Walter de Gruyter.

Rijkhoff, J. (2007). Word classes. *Language and Linguistics Compass 1, 709–726.*

Vogel, P. M., Comrie, B. (Eds.) (2000). *Approaches to the Typology of Word Classes* (Empirical Approaches to Language Typology 23). Berlin/New York: Mouton de Gruyter.

Vulanović, R. (2003). Grammar efficiency and complexity. *Grammars 6, 127–144.*

Vulanović, R. (2007). On measuring language complexity as relative to the conveyed linguistic information. *SKY Journal of Linguistics 20, 399-427.*

Vulanović, R. (2008). A mathematical analysis of parts-of-speech systems. *Glottometrics 17, 51-65.*

Vulanović, R. (2009). Efficiency of flexible parts-of-speech systems. In: Köhler, R. (Ed.), *Issues in Quantitative Linguistics*: *136-157*. (= Studies in Quantitative Linguistics 5) Lüdenscheid: RAM-Verlag.

Vulanović, R. (2010). Word order, marking, and a two-dimensional classification of parts-of-speech system types. *Journal of Quantitative Linguistics 17, 229-252.*

Vulanović, R., Baayen, H. (2007). Fitting the development of periphrastic *do* in all sentence types. In: Grzybek, P., Köhler, R. (Eds.), *Exact Methods in the Study of Language and Text*: *679-688*. Berlin-New York: Mouton de Gruyter.

Vulanović, R., Köhler, R. (2009). Word order, marking, and parts-of-speech systems. *Journal of Quantitative Linguistics 16, 289-306.*

Vulanović, R., Miller, B. (2010). Grammar efficiency of parts-of-speech systems. *Glottotheory 3/2, 65-80.*

Adresses of authors and editors

Altmann, Gabriel
Stüttinghauser Ringstr. 44
D-58515 Lüdenscheid
Germany
E-mail: RAM-Verlag@t-online.de

Best, Karl-Heinz
Im Siebigsfeld 17
D-37115 Duderstadt
Germany
E-mail: kbest@gwdg.de

Čech, Radek
Domluvilova 231/9
CZ-75701 Valašské Meziříčí
Czech Republic
E-mail: radek.cech@osu.cz

Drebet, Viktor
Lehrstuhl für Deutsch
Nationale pädagogische
Hnatjuk-Universität
Krywonosastr., 2
UA-46027 Ternopil
Ukraine
E-Mail: vdrebet@yandex.ru

Inamdar, Atul S.
Durga Apartment No.1, Flat 6,
Raghuvirnagar, Opp.S.F.S.High
School,
Jalna Road Aurangabad
Maharastra
India
Email: atulchi@yahoo.com

Kantemir, Sergej
Lehrstuhl für germanische,
allgemeine und vergleichende
Sprachwissenschaft
Jurij-Fedjkowytsch-
Nationaluniversität Tscherniwzi
Kozjubynskoho-Str., 2
UA-58000 Tscherniwzi
Ukraine
E-mail: sergej.kantemir@gmail.com

Kelih, Emmerich
Institut für Slawistik
Universität Graz
Merangasse 70/1
A-8010 Graz
Austria
E-mail: emmerich.kelih@uni-graz.at

Levitskij, Victor V.
Lehrstuhl für germanische,
allgemeine und vergleichende
Sprachwissenschaft
Jurij-Fedjkowytsch-
Nationaluniversität Tscherniwzi
Kozjubynskoho-Str., 2
UA-58000 Tscherniwzi
Ukraine
E-mail: vlevizky@gmail.com

Lvova, Nadija
English Language Department
Chernivtsi National Yurij Fedkovych
University
2 Kotsiubynskyy Street
UA-58000 Chernivtsi
Ukraine
E-mail: nadezhdalvova@yahoo.com

Mačutek, Ján
Institut für Slawistik,
Karl-Franzens Universität Graz
Merangasse 70
A-8010 Graz
Austria
E-mail: jmacutek@yahoo.com

Matskulyak, Yuliya
Lehrstuhl für germanische,
allgemeine und vergleichende
Sprachwissenschaft
Jurij-Fedjkowytsch-
Nationaluniversität Tscherniwzi
Kozjubynskoho-Str., 2
UA-58000 Tscherniwzi
E-mail:
yuliya.matskulyak@gmail.com

Milička, Jíři
Charles University in Prague
Faculty of Arts
Institute of Comparative Linguistics
E-mail: Milicka@centrum.cz
www.milickacz.

Popescu, Ioan-Iovitz
Str. Fizicienilor Nr. 6, Bloc M4
RO-077125 Magurele/Ilfov
Romania
E-mail: iovitzu@gmail.com

Prabhu-Ajgaonkar, S.G.
Contact: see under Inamdar

Riley, Charles
Sterling Memorial Library
Yale University
New Haven
Connecticut, USA
Contact: see under Rovenchak

Rovenchak, Andrij
Department for Theoretical Physics,
Ivan Franko National University of
Lviv
12 Drahomanov Street
UA-79005 Lviv
Ukraine
E-mail: andrij@ktf.franko.lviv.ua;
andrij.rovenchak@gmail.com

Sherman, Tombekai
Abidjan, Côte d'Ivoire; Athinkra
LLC. Contact: see under Rovenchak

Soloviova, Olga
English Language Department
Chernivtsi National Yurij Fedkovych
University
2 Kotsiubynskyy Street
UA-58000 Chernivtsi
E-mail: poloskotun@gmail.com

Vulanović, Relja
Department of Mathematical
Sciences,
Kent State University at Stark
6000 Frank Ave NW, North Canton
Ohio 44720, USA.
E-mail: rvulanov@kent.edu