Book review. *Kvantitativnaja Lingvistika: Issledovanija i modeli* (Klim-2005). Materialy Vserossijskoj naučnoj konferencii (6-10 ijunja 2005 g.). Novosibirsk: Novosibirskij Gosudarstvennyj Pedagogičeskij Universitet. (Redakcionnaja kollegija A.A. Polikarpov, G.G. Sil'nickij, V.V. Poddubnyj). Reviewed by **Emmerich Kelih.**

0.

The book under review, "Quantitative linguistics: Analyses and models (Klim 2005). Proceedings of the All-Russian Scientific Conference of June 6-10, 2005" (hereafter Klim 2005), is an omnibus volume containing the proceedings of the conference on quantitative linguistics held in Novosibirsk (Russia).

The conference was organized by the National Novosibirsk Pedagogical University – the publisher of the omnibus volume – under the chairmanship of the scientific secretary of the conference Victor V. Kromer. The lasting importance of the conference is attested to by the fact that 45 scientists[1] took part in it. The volume, which appeared in 2005, provides an ideal starting point for reviewing individual contributions.

The 370-page volume (Klim 2005) contains 35 contributions, written in Russian by 39 authors. The editors ("redakcionnaja kollegija") were A.A. Polikarpov, G.G. Sil'nickij and V.V. Poddubnyj. Contrary to usual practice, Klim (2005) does not – unfortunately – contain either a preface or an introduction by the editors. This would provide at least an initial survey of the main topics of the conference or the published papers. Also somewhat striking is that the book is not subdivided into thematic sections. It would, of course, be possible to recognize a thematic level-subdivision on the basis of the order of the papers (from phonosemantics through style studies up to general theoretical questions); but in this review we shall, rather, consider the papers as being classified into three domains: (1) quantitative text linguistics, (2) quantitative linguistics, and (3) computational linguistics/data mining. The review of particular analyses will be followed by (4) a general summary of the most important research areas in Russian quantitative text and language analysis resulting from Klim (2005).

1.

The first thematically extensive division in Klim (2005) is quantitative text linguistics. The defining factor of this systematisation is an explicit interest in individual texts (distinguishing style and verse analysis respectively), authorship attribution, and general considerations on "text homogeneity".

Let us start with the domain of quantitative style analysis, which rests quite on a long-standing tradition in quantitative research in text science. A consistent interest in the study of sentence length and the frequency of syntactic constructions may be observed in three of the contributions to Klim (2005).

T.V. Džurjuk ("Sentence length as a parameter of individual author style of German writers at the beginning of the 20th century: based on data from F. Kafka, I. Keun, Th. Mann and K. Tucholsky", pp. 182-194)[2] reports on studies of length and frequency of punctuation in sentences. The frequency of sentence lengths has been examined statistically in more detail by means of the chi-square contingency test and the contingency coefficient. According to the author, sentence lengths pooled to groups ("short", "mean" and "long" sentences) give important information about the stylistic individuality of the writers mentioned above. The high proportion of short sentences found in the texts motivates the examination of the hitherto neglected text-internal heterogeneity: we may distinguish direct speech, speech insertions and narrative text parts. Even if the examination of these text components is restricted to the state-

[1] The number of participants has been calculated from the conference's homepage (http://klim.nightmail.ru/).
[2] The translations of individual titles were made by the reviewer. In some cases the titles have been shortened on stylistic grounds without substantially changing the meaning.

ment of absolute frequencies for each writer, it would clearly be possible to derive from these facts a broader relationship between sentence length and its affiliation to a certain text component.

Quite similar, with regard to the methods applied, is the contribution of *Ju. P. Bojko* ("Selectivity of the author's style at the sentence level", pp. 303-315). The author is not directly interested in sentence length, but rather in the frequency of syntactic constructions (subject, predicative clauses, etc.) in English writers. Furthermore, in this case the ascertained "range" of frequency of occurrence is interpreted as originating from the individuality of the author. Because of the fulfilled syntactic segmentation of the sentence, its length should be examined as a subsequent step.

Apart from this descriptive capturing of authorial style based on subtly distinguished units of investigation (parts of speech, clause) and tested by means of the chi-square criterion, *V.V. Poddubnyj* and *O.G. Ševelev* ("Comparison and cluster analysis of text features (frequency) on the basis of the hypergeometric criterion", pp. 205-217) address the use of the clustering procedure in quantitative text analysis and text typology. To this end, the authors use the sentence length in works by several Russian authors. The text size is given in the somewhat unusual specification of megabytes (MB). The authors examine in greater detail whether the combination of the assumed "syntactic" parameters, e.g. frequency of sentences (measured in quite arbitrary fixed lengths of 5, 10 etc. words) or the proportion of chapters, can influence the "effectivity" of text classification when using the clustering method. The partially ascertained differences between individual clustering results are, however, not interpreted further; the authors are content with the resulting classification. Hence, on this point we must agree with the authors' statement that further experimental examinations and a theoretical foundation of the results are required (cf. Klim 2005: 216).

Three further articles are dedicated to the quantitative study of disputed authorship. The well-known St Petersburg specialist in this domain, *M.A. Marusenko*, discusses jointly with *E.E. Mel'nikova* and *E.S. Rodionova* ("Attribution of anonymous and pseudoanonymous articles published in the journals 'Vremja' and 'Epocha' in 1861-1865", pp. 283-293) the problem of certain articles of disputed authorship that have, up until now, generally been ascribed to Dostoevsky. The quantitative criteria for the solution of this problem are the length of simple sentences, the number of sentences without a subject, and some further morphosyntactic considerations. With regard to the categorization of the applied properties, Marusenko (1990) is still very helpful. The previous attribution of these articles to Dostoevsky is challenged on the basis of computed Euclidean distances. Instead, it is supposed that the disputed articles could have been written by several authors.

The same problem is followed up in the article by *N.V. Semjannikova* ("Verification of attribution methods with translated texts", pp. 294-302). However, the author is less interested in content problems than in testing the significance of different methods. She postulates the inadequacy of (1) the Q-sum method (cf. Tweedie 2005: 393f.) for Russian, although no empirical evidence is presented. On the other hand, (2) N.A. Morozov's well-known method of "authors' invariant" (frequency of prepositions) clearly allows, in the opinion of the author, statements about the author's individuality to be made. This interpretation arises solely from graphical representations and is not – as should be expected – substantiated by statistical testing. The greatest success rate is ascribed to (3) the method of frequency of letter bigrams, based on Khmelev's (2000) considerations. This contribution should be considered an approximation to the problem of finding an adequate method of attribution. The stated intention, namely comparing translated texts by different writers, may be considered quite a prolific perspective.

A further methodologically oriented contribution, which nevertheless can be placed in the domain of authorship attribution, is that of *A.P. Kovalevskij* ("Application of the invariance

principle in the analysis of text homogeneity", pp. 195-204).[3] The author is interested in the problem of whether a text is homogeneous, i.e. not written by two or more authors. The concrete motivation is the discovery of an objective criterion for identifying plagiarism, for example in schoolwork, which do not necessarily originate from one author. The approach taken here to this interesting problem is to some extent original. In ascertaining text homogeneity or individual author style, the starting assumption is that the authors use the same words with different frequency. In order to obtain an adequate basis for investigation, the author sets up frequency lists of words for several Russian novels (by Dostoevsky, Tolstoy, and Bulgakov). From these frequency lists, an intersection set of words occurring in all the novels is constructed. This intersection set is called "standard language" *Z* and it contains 27,000 word forms. Using a special stochastic process (Brownian motion), the author tries to show that a sudden increase in the occurrence of frequent words of the "standard language" in a text block (which is not clearly defined) is a break of homogeneity and signals plagiarism. The relevance of the method is illustrated by a comparison of a literary text, its retelling by a pupil, and a combination of these two texts. In this paper, two things seem to be lacking. On the one hand, the reader feels the need for a discussion of the complex and multi-layered problem of text plagiarism; on the other hand, we might ask for the linguistic reasons for the relevance of high-frequency words in plagiarism detection. Further, it is quite conspicuous that the rich Russian tradition of systematic compilation of frequency dictionaries is totally ignored. These could surely serve as a starting point for this approach. Last but not least, the law of Frumkina must be consulted as an initial step.

The contribution of *O.N. Grinbaum* ("Quantitative analysis of the verse: from linguistics to 'art-metrics'", pp. 94-107) is concerned with quantitative verse analysis. The theoretical interest is concentrated on rhythm, which together with meter represents an important research field both in literary and linguistic domains. Grinbaum is, however, not interested in the empirical study of rhythm, but rather in an adequate definition of this phenomenon. In this connection, he provides a thorough survey of one hundred years of the Russian discussion of the concept of "rhythm". From this study, it is evident that he does not considers rhythm in the structuralist sense as a sequence of accents, but defines it as a "harmony of relationships". In addition, such a conception, in sense of the Golden Mean, can be quantified. The author follows this up intensively elsewhere (Grinbaum 2000)

A.S. Gumenjuk and *A.S. Kostyšin* work directly in the domain of quantitative verse analysis. In their investigation ("Acoustic elements of verses and the formal procedures of their recognition", pp. 34-43), the authors present a new procedure of text segmentation, whose fundamentals have been shown in detail already in Gumenyuk, Kostyshin and Simonova (2002) and Gumenjuk et al. (2004). A text is segmented into acoustic text units called 'consonances' which can be set up in different length forms on the basis of neighbouring phonemes with the aid of an algorithm.

Actual empirical examination of these acoustic text units is given in the next contribution by *A.S. Kostyšin* ("Investigations into a segmentation algorithm", pp. 44-60). Here texts are segmented by means of the algorithm mentioned above in acoustic text units of different length, and a frequency dictionary is compiled. In this connection, the well known hypothesis of Ju.K. Orlov concerning the validity of Zipf-Mandelbrot law for full and closed texts (Zipf size) is applied as a criterion or test procedure. It should corroborate the correctness of the chosen segmentation of poetic texts in its psycholinguistic relevance.

Some further articles in Klim (2005) can be classified as selected problems in quantitative text analysis and empirical literary science. *N.L. Zeljanskaja* ("Philological reception of a literary fact: experimental analysis and modelling", pp. 61-72) presents an empirical inquiry

[3] A reference from this contribution could not be stated correctly, namely "Herdan, E. Calculus of legomena / E. Herdan. – N.-Y. 1964" (cf. Klim 2005: 204).

into the assessment of the literary and cultural-historical relevance of classical Russian authors. To this end, she interviews test persons considered by her to be professional readers (philologists, literature scientists). Unfortunately, it is not possible to reconstruct from this article what is to be examined. Even if there are cursory hints to the statistical evaluation of the questionnaires – up to now 7 persons have been interviewed – the aim and the method of this contribution remains generally unclear.

An innovative and original contribution to experimental methods in the domain of linguistics and text science goes back to *K.I. Belousov* ("Modelling the interplay of intratextual spaces", pp. 73-93): 21 test persons segmented a given text into "micro-themes" and on lexico-semantic basis ascribed the word occurring therein to another word. The frequency of this ascription performed by test persons serves as a starting point for a graphic representation of the so-called semantic-intratextual relationships. In the next step, these subjective text relationships are supplemented by further prosodic properties: the test persons must expressively read aloud the text they have segmented into semantic relations. The reading is recorded. In the opinion of the author, it is possible to set up a relationship of "semantic" and "prosodic" contour between the semantic-lexical intratextual relations of a text and the loudness of the declaimed part of the text. Apart from the unclear meaning of "intratextal" semantic text relationships, one might surely question whether loudness, as a property of the prosodic contour taken into account by the author, should not be supplemented by further experimental phonetic factors (intonation, speech pause, etc.)

The investigation of *Ju.N. Kovšova* and *V.A. Seleznev* ("Measurement of the activity of the supporting and the logical language functions for the ascertainment of text quality", pp. 137-145) treats the problem of a quantifiable evaluation of the quality of mathematics textbooks. The starting point is text parts with illustrating ("supporting") and explanative ("logical") functions. Somewhat strange is the quantitative determination of these text parts: it is not the frequency or the length of the passages having this function in textbooks but their *area*, measured in square centimetres. The areas yield some ratios but these are not standardized. Their values in individual chapters are simply juxtaposed with those of other textbooks. This method, which according to the authors should enable a statement about the "didactic" quality of textbooks, must be critically scrutinized because it represents a very raw, exclusively optical approximation to a very interesting problem.

A slightly different, but in any case acceptable, view of text homogeneity can be found in *V.A. Seleznëv* and *E.V. Isaeva* ("The Hurst parameter in word sequences", pp. 146-151). Considering the text as a time series, the sequential order of word lengths measured in terms of letter numbers is analysed: first in full, closed Russian novels, then in texts which were "shuttled" by means of a special algorithm. As a matter of fact, Hurst's parameter is different in differently "processed" texts. But also, in this case, the ascertained differences are interpreted exclusively on the background of graphical presentations. Nevertheless, the method raises the possibility of quantitatively scrutinizing the semantic "unity" of texts.[4]

Similar to this contribution is a further examination of text homogeneity by means of the Hurst parameter and Brownian motion by *N.S. Zakrevskaja* ("Study of text homogeneity using the moving average", pp. 26-33). The author studies the sequence of word lengths, measured in terms of syllable numbers, in two literary texts and in a "quasi-text" representing a cumulative mixture of the two texts. In contrast to the previous article, the author is slightly more cautious in the interpretation of her results, and understands her contribution explicitly as a test of a method using texts as illustrations (cf. Klim 2005: 33).

[4] For this investigation one would expect at least a reference to the first discussion of Hurst´s parameter by Krylov (1994: 113f) or the detailed discussion accompanied by a series of empirical investigations concerning sentence length by Hřebíček (1997: 132ff.).

2.

Another large thematic block in Klim (2005) consists of works belonging to the "core domain" of quantitative linguistics. Here, the characterisation of the text/author is of secondary importance. Rather general language properties and language phenomena that are scrutinized by means of quantitative methods become the focus of attention.

Four contributions of the "Smolenks Group" led by Prof. V.V. Silnickij belong in this block. The starting point of this group is the use of different correlation methods for stating the mutual relationships between phonetic, morphosyntactic and semantic features of language. Until now, the English verb has stood especially in the focus of attention (cf. Sil'nickij et al 1990). On the basis of the articles published in Klim (2005), some new meaningful problems and methodological innovations can be envisioned.

In the contribution of *A.G. Sil'nickij* ("Quantitative semantic classification of "economic" situations of modeling English verbs", pp. 167-181) 500 English verbs (especially those describing economic relations between subjects) are analyzed as to their syntactic-semantic properties. On the basis of a maximal number of different criteria using correlation analysis, verb subclasses are identified that should be put in mutual relations. A further step using cluster analysis corroborates the result.

A classification of German verbs can be found in *E.A. Il'jušina und D.S. Goršenin* ("Testing the feature system of German verbs using multivariate procedures", pp. 364-368). Here a classification of 4892 verbs is performed using 33 properties (phonetic to semantic) and the clustering method. Because of the shortness of the paper, and the omission of tabular and graphical presentations of individual steps in the analysis, no direct conclusions may be drawn for the time being.

In contrast to this, the aim of *L.A. Kuz'min's* contribution ("Correlations between different levels of adjectives in modern English", pp. 108-121) is much more clearly evident: he is concerned with internal mutual relations within the class of English adjectives and connects phonetic, morphological, morphosyntactic, etymological, chronological and semantic properties, testing statistically the significance of interrelations.

The contribution by *G.G. Sil'nickij* ("Correlation and discriminance analysis of languages and language properties", pp. 152-166) does not concern language internal correlations, but rather a multivariate discriminance analysis of typological properties. In 78 areally and genetically different languages, first all correlated features (out of 47 phonetic and grammatical ones) are excluded. The non-correlated features are the starting point of discriminance analysis, yielding no corroboration of the known genetic classification, but five new typological groups. All in all, this is an interesting paper which should be further discussed. It must be remarked that correlation and discriminance analysis as well as factor analysis are merely inductive explorative means not directly yielding theoretical results. Classification itself is a very shaky ground, easily manipulable. It is theoretically prolific if it can be derived from a theory. Hence its direct relevance is not evident. Nevertheless, this direction at least gives rise to many hypotheses which can later on be founded theoretically.

Semantic structures and associated partial domains are explicitly studied in three contributions. *A.A. Vengrenovič* and *V.V. Levickij* ("Quantitative parameters of synonymy in German", pp. 228-231) study the noun in German quantitatively. An analysis of three synonymy dictionaries (containing 64,076 nouns) displays among other things a significant negative correlation between the synonyms of a lemma and the frequency of stylistic markers. No less interesting is the discovery of a relation between the gender category of nouns and the number of synonyms.

The contribution of *L.V. Gikov* and *G.V. Gikova* ("Adverb-verb connections in German", pp. 232-243) studies the frequency of word class combinations on the basis of random

samples from prose texts by German writers. The complete sample contains 3,000 adverb-verb connections. In this corpus, the adverbs are divided into six subclasses (temporal local, modal, causal, conditional and consecutive) and the verbs are classified according to three groups (state, process and activity verbs), consisting of a further 26 subcategories. In this set of data, the authors try to ascertain whether certain adverb-verb combinations, having given semantic-syntactic properties, are significantly frequent. The authors succeed in filtering out about 20 very frequent combinations using the chi-square test, and give relevant information about the strength of adverb-verb combinability.

The investigation of *N.L. L'vova* ("Study of phonosemantic interrelations between consonant clusters at the beginning of words in different texts", pp. 3-10) concerns the question of whether consonantal bigrams at word beginnings have some relation to the functional style of the text. A similar approach can be found in Lvova (2005). The author shows that in English poetic, literary and journalistic texts, the frequency of some consonant clusters possesses a functional stylistic property. This result, tested by means of the chi-square test, should also be tested in larger samples (so far the results hold for 12 poems, 8 literary texts and 10 press texts, where the sample size is never greater than 500 words). In that case it could, perhaps, be corroborated that the frequency of consonant clusters at the beginning of words is not a general language-specific phenomenon but, as a matter of fact, a genre-specific feature. Here, perhaps, more references concerning the study of iconism might be expected.

A.A. Polikarpov dedicates his paper ("Evolutionary foundations of Menzerath's law and the search for the dependence of morpheme length on its positional features", pp. 351-363) to Menzerath's law, which is well established in quantitative linguistics. This article (see also Polikarpov 2006) tries to integrate Menzerath's law into the theory of "life cycle" of words developed by the author. While (traditionally) Menzerath's law describes the relation between the size of linguistic constructs and the length of their constituents, the author tries to supplement it with a positional aspect. He concentrates on the length of suffixes and prefixes in dependence on the distance (position) from their stem morphemes. As a matter of fact, it can be observed that prefixes which are more distant from the stem morpheme tend to be longer, while suffixes get smaller with increasing distance from the "centre". So far the oinly data at our disposal concerns Russian. In a dictionary of Russian word forms segmented morphologically compiled by the author (50,747 items), morpheme length (measured in the number of letters) is considered separately for prefixes, roots and suffixes. Now, if we consider the length of prefixes in dependence on their distance from the stem (i.e. whether a word form contains one, two or three prefixes) we see that with increasing distance the prefixes get longer (an example from the published data: prefixes in the third position to the left of the stem morpheme have a mean length of 2.597 letters, those in the second position 2.249 and those in the first left position 2.08 letters). On the other hand, with increasing distance from the stem morpheme the length of suffixes decreases. This tendency is not as markedly expressed as the dependence of prefix length on its position; a kind of length oscillation can be observed. For the time being there are only data from the morphologically very rich Russian, but they are interesting solely because of their quantitative ratios: in the given corpus, word forms have maximally 3 prefixes and 7 affixes. A comparison with other languages would show that this length tendency is a special feature of Russian or even Indo-European languages; it is not present in strongly agglutinating languages.

In this contribution, relationships between morpheme length and word age, frequency, and semantic class are postulated. These interesting cross-connections represent the core part of the "life cycle" theory which will surely be discussed intensively in (quantitative) linguistics. Above all the integration of the positional dependence of morpheme length into the "traditional" Menzerath's law could be of great interest.

There is only one contribution in Klim (2005) to glottochronology, which is very intensively studied in Russian linguistics, namely that by *L.A. Selezneva-Eleckaja* ("Semantic factors of different susceptibility levels against decay of units in the glottochronological list: using data from Indo-European languages", pp. 322-335), in which the author tries to classify the lexical items of the Swadesh list and their modifications according to semantic fields. The aim of such an investigation is to ascertain whether there is a relation between the semantic content and the degree of survival of lexical units. According to the author, this holds for certain groups (e.g. words expressing feelings), but this can in turn be understood as a hint at the internal non-homogeneity of the word forms in the Swadesh lists. Here we should recall V.V. Levickij's objection that the classification of semantic units e.g. according to the degree of abstractness etc. is notoriously ambiguous (cf. Levickij 2005: 462). In addition to this global problem there is another global flaw in this work: the author operates exclusively with percentages and graphical presentations. Such a procedure in no case meets the methodological expectation of quantitative linguistics which approaches data of this kind with at least a statistical test.

The contribution by *M.V. Usmanova* ("Linguistic and psychological peculiarities of gender: quantitative aspects", pp. 316-321) concerns gender study. The author performs an empirical enquiry (48 test persons) into gender-specific language customs in Russian. On the basis of the evaluation of conversations about certain themes, she could ascertain differences in the use of modal verbs between men and women.

The article of *M.K. Timofeeva* ("Measurement and modelling techniques of the naturalness level of systems with natural language interaction", pp. 336-350) is of a rather theoretical nature. The author shows the possibilities and limits of a quantitative differencing of natural and artificial languages which, according to the author, will grow in importance in the man-machine dialogue.

3.

The third and last block of papers in Klim (2005) address general aspects of computer-based processing of language texts, corpus linguistics and "data mining". Statistical evaluations are rare here but we report on them briefly, in order to round off the depiction of Klim (2005).

Surprisingly there are only two articles in the volume that can be attributed to corpus linguistics. While *I.V. Arzamasceva* ("Statistical investigation of German terminology to the theme 'Fuzzy logic' by means of the program Fuzzy-Base", pp. 244-255) is concerned with the establishment of a data bank (frequency dictionary) of the German lexicon in the domain of fuzzy sets, *A.I. Izotov* ("Quantitative aspects of the description of functional-semantic categories of the imperative in present-day Czech", pp. 122-136) investigates the frequencies of Czech imperative constructions on the basis of the Czech National Corpus. The data are contrasted with those of Russian.

V. G. Klimov ("The linguistic component of the computer based processing of data from natural languages: problems and perspectives of information-communicative technologies in school education", pp. 11-25) dedicates his article rather to general problems of computational linguistics than to the use of quantitative methods. The point is a basic discussion of automatic semantic text analysis. Also the contribution by *A.M. Naletov* ("A computer-based system for the analysis of natural language texts", pp. 218-227) is oriented to basic research and concerns the analysis of semantic networks, a theme that belongs rather to the domain of "small worlds".

An attempt at the automatic recognition of "similar" texts can be found in the article by *E.N. Benderskaja* and *S.V. Žukova* ("Processing of text information by means of chaotic neuronal networks", pp. 271-282) who discuss some concepts of fuzzy logic. Finally, *V.D.*

Gusev and *N.V. Salomatina* ("L-gram analysis of natural language texts and its possibilities", pp. 256-270) present an algorithm for the identification of steady collocations (called here L-grams) and supply the first frequencies. Even if the results are not further evaluated statistically, it is at least an important step toward quantitative analysis on the syntactic level.

4. Summary

Having reported on 31 contributions in Klim (2005), encapsulating a wide and heterogeneous spectrum of applications of quantitative methods in the Russian linguistics and text science, it is possible to detect some general tendencies. From the great number and wide scope of the articles, which as far as content and methodology are concerned display a high level of mathematical-statistical competence, we may draw the conclusion that this omnibus volume can be considered representative[5] of the present state of the art in Russia.

The individual contributions provide an excellent insight into some special domains developed in recent years: worth mentioning are the correlation work from Smolensk, the St Petersburg analyses of authorship attribution and verse analysis, the ideas on the "word life cycle" and an explicit interest in quantitative investigations of semantic structures (synonymy, polysemy, etc.). This is the profile of the Russian quantitative linguistics and text science.

On one hand this profile is a conscious continuation of the experience of Russian quantitative linguistics. Worth mentioning is especially Ju. K. Orlov's et al. hypothesis of "Zipf size" which provides even today a starting point for other innovative investigations. On the other hand, we may recognize a break in tradition. In Klim (2005) there are neither references to the works of Ju.A. Tuldava and Ju.A. Krylov, which are still current, nor to the group "Statistika reči« which dominated the field of statistical investigations in Russia for decades.

The omission of these references may be ascribed to the dynamics or to the sense of a new era in Russian scientific enterprise. At least from the point of view of the application of quantitative methods, this process brings forth a series of new, partially innovative and original approaches. Though innovation and originality is welcome in every scientific discipline, there must be an equal emphasis on the maintenance of certain "standards". "Standards" here are meant in the sense of the "working method of quantitative linguistics" (cf. Altmann 1972: 3ff.; Köhler 2005: 8ff.)[6] which controls the course of linguistic or text-analytical investigations using quantitative methods, and which is not prescriptive but can, nevertheless, be considered a raw and acceptable guideline.

---

[5] Klim (2005) is surely comparable with the 1991 conference in Smolensk called "Ėvrističeskie vozmožnosti kvantitativnych metodov issledovanija jazyka//Heuristic possibilities of quantitative methods in language analysis (cf. Sil'nickij, Tuldava and Polikarpov 1991) and with "2-aja Meždunarodnaja konferencija po kvantitativnoj lingvistike//Second international conference on quantitative linguistics" (cf. Polikarpov 1994) in Moscow, September 1994. The proceedings of these conferences contain only longer abstracts.

[6] In this connection we expressly refer to the recently published handbook of quantitative linguistics (cf. Köhler, Altmann and Piotrowski 2005). This book represents not only the actual state of the arts in quantitative linguistics concerning its theoretical and methodological bases but also a compact source of information allowing to embed one own research in a wider framework. In other words, Klim (2005) displays a partly small readiness to discuss and absorb work from the non-Russian area. This is a phenomenon already criticized by the well-known Russian verse theoretician B.V. Tomaševskij who laid the central foundation stones of the statistical and probabilistic verse analysis and discovered this feature in the Russian verse statistics in the second and third decade of the 20th century. Especially the "domoroščennaja statistika/homemade statistics" was a thorn in his side. Tomaševskij emphasized as an alternative the »philological statistics« which should explicitly take note of and discuss works beyond the Russian scientific area (cf. Tomaševskij 1923: 139).

The formulation of a linguistic/text-analytical hypothesis, the operationalization of the units to be quantified, the translation of the hypothesis in the language of statistics, and above all the (not always easy) choice of an adequate statistical method, should as a rule precede an interpretation. It is not our aim to evaluate the individual contributions according to this criterion. Rather we wish to say that linguistic and text-analytic investigations claiming to work "quantitatively" should not in general be based only on graphical presentations of characteristics or ratios. They should, in agreement with the required intersubjectivity, follow the above-mentioned course of work in order to present a lasting and relevant contribution.

Apart from this partially noticeable methodological weakness, this extensive volume is, nevertheless, a successful and exciting mixture of tradition, innovation and originality. It is to be hoped that this remains characteristic of quantitative procedures in linguistics and text analysis in the future, not only in Russia.

References

Altmann, G. (1972): Status und Ziele der quantitativen Linguistik. In: Jäger, S. (ed.), *Linguistik und Statistik*. Braunschweig: Vieweg, 1-9.

Dshurjuk, T.V.; Levickij, V.V. (2003): "Satztypen und Satzlängen im Funktional- und Autorenstil", in: *Glottometrics* 6, 2003, 40-51.

Grinbaum, O.N. (2000): *Garmonija strofičeskogo ritma v ėstetiko-formal'nom izmerenii.* Sankt Peterburg: Izdatel'stvo Sankt-Peterburgskogo Universiteta.

Gumenjuk, A.; Kostyshin, A.; Borisov, K.; Salnikova, O. (2004): "On the acoustic elements of a poem and the formal procedures of their segmentation", in: *Glottometrics* 8, 42-67.

Gumenyuk, A.; Kostyshin, A.; Simonova, S. (2002): "An approach to the analysis of text structure," in: *Glottometrics 3,* 61-89.

Hřebíček, L. (1997): *Lectures on Text Theory.* Prague: Oriental Institute, Academy of Sciences of the Czech Republic.

Khmelev, D. (2000): "Disputed Authorship Resolution through Using Relative Empirical Entropy for Markov Chains of Letters in Human Language Text", in: *Journal of Quantitative Linguistics*, 7, 3; 201-207.

Köhler, R. (2005): Gegenstand und Arbeitsweise der Quantitativen Linguistik. In: Köhler, R.; Altmann, G.; Piotrowski, R.G. (eds.): *Quantitative Linguistik/Quantitative Linguistics.* Berlin u.a.: de Gruyter, 1-16. [= Handbücher zur Sprach- und Kommunikationswissenschaft, 27].

Krylov, Y.K. (1994): Hurst's law as a Universal Law of Quantitative Linguistics of a coherent text. In: Polikarpov, A.A. (ed.) (1994), 113-114.

Levickij, V. (2005): Polysemie. In: Köhler, R.; Altmann, G.; Piotrowski, R.G. (eds.): *Quantitative Linguistik/Quantitative Linguistics.* Berlin u.a.: de Gruyter, 458-464. [= Handbücher zur Sprach- und Kommunikationswissenschaft, 27]

Lvova, N.L. (2005): "Semantic functions of English initial consonant clusters", in: *Glottometrics* 9, 21-28.

Marusenko, M.A. (1990): *Atribucija anonimnych i psevdoanonimnych literarturnych proizvedenij metodami teorii raspoznavanija obrazov.* Leningrad: Izdatel'stvo Leningradskogo Universiteta.

Polikarpov, A.A. (ed.) (1994a): *Qualico-94. 2nd International Conference of Quantitative Linguistics. September 20-24 1994//2-aja Meždunarodnaja konferencija po kvantitativnoj lingvistike 20-24 sentjabrja 1994 goda.* Moskva: MGU im. M.V. Lomonosova, Filologičeskij fakul'tet.

Polikarpov, A.A. (2006): Towards the Foundations of Menzerath's Law. In: Grzybek, P. (ed.): *Contributions to the Science of Language.* New York u.a.: Springer, 255-272.

Sil'nickij G.G.; Andreev, S.N; Kuz'min, L.A.; Kuskov, M.I. (1990): *Sootnešenie glagol'nych priznakov različnych urovnej v anglijskom jazyke.* Minsk: Navuka i Tėchnika.

Sil'nickij, G.G.; Tuldava, Ju.A.; Polikarpov, A.A. (eds.) (1991): *Ėvrističeskie vozmožnosti kvantitativnych metodov issledovanija jazyka. Tezisy dokladov Vsesojuznogo seminara v gor. Smolensk 11-13 sentjabrja 1991 g.* Smolensk: Smolenskij SGPI.

Tomaševskij, B.V. (1923): "Problema stichotvornogo ritma" in: *Literaturnaja mysl'* 2, 124-140.

Tweedie, F.J. (2005): Statistical models in stylistics and forensic linguistics. In: Köhler, R.; Altmann, G.; Piotrowski, R.G. (eds.): *Quantitative Linguistik/Quantitative Linguistics.* Berlin u.a.: de Gruyter, 387-397. [= Handbücher zur Sprach- und Kommunikationswissenschaft, 27].

Rezension: Gabriel Altmann/Viktor Levickij/Valentina Perebyinis (Hg.): *Problemy kvantytatyvnoji lingvistyky: zbirnyk naukovych pracj* (Probleme der Quantitativen Linguistik: Sammelband). Černivci: Ruta, 2005. 352 Seiten. Von **Juri Kijko.**

Der vorliegende Band mit internationalen Beiträgen präsentiert die neuesten Ergebnisse der Quantitativen Linguistik als einer wissenschaftlichen Disziplin und verfolgt das Ziel, ihren heutigen Stand und ihre Perspektiven darzustellen. Der Sammelband besteht aus fünf Teilen: 1. Quantitative Linguistik als eine wissenschaftliche Disziplin, 2. Zur Anwendung quantitativer Methoden in der Linguistik, 3. Quantitative Untersuchungen der Sprach- und Texteinheiten, 4. Korpuslinguistik, 5. Quantitative Gesetze, Verteilungen und Programme. Der Band vereint 22 Beträge von Sprachwissenschaftlern aus acht Ländern.

Am Anfang des ersten Teils steht als Vorwort ein Beitrag eines der Herausgeber Gabriel Altmann (Deutschland), der seine Überlegungen zu Mode und Wahrheit in der Wissenschaft anstellt und für die Quantitative Linguistik plädiert.

Auf die Ziele und Methoden der Quantitativen Linguistik gehen Reinhard Köhler und Gabriel Altmann (Deutschland) in ihrem Beitrag (S. 12-41) ein. Die Autoren begründen die Einführung von quantitativen Konzepten, Modellen und Methoden in die Linguistik sowie in die Textwissenschaft, wie es in den betreffenden Disziplinen der Fall ist.

Im einem weiteren Beitrag (S. 42-59) von Gabriel Altmann und Peter Meyer (Deutschland) setzen die Autoren Überlegungen über die Quantitative Linguistik und ihre Rolle in der Sprachwissenschaft fort. Sie sind der Meinung, dass „the intervention of physicists in linguistics can help us to open our science for the theory of general systems".

Ramon Ferrer i Cancho (Italien/Spanien) wendet sich (S. 60-75) der Erforschung der Struktur des syntaktischen Netzes zu und entdeckt aufgrund der neuesten Ergebnisse mögliche Wege zum Verstehen der universalen Eigenschaften der menschlichen Sprache.

Karl-Heinz Best (Deutschland) stellt in seinem Beitrag (S. 76-88) fest, „dass die Quantitative Linguistik in den deutschsprachigen Ländern zur Zeit einigermaßen prosperiert, verdankt sie ganz wesentlich den vielfältigen Anregungen der osteuropäischen Forscher und den verbesserten Kontaktmöglichkeiten, die sich seit etwa 1990 entwickelt haben".

Der zweite Teil des Bandes thematisiert die Erfahrungen der Wissenschaftler bei der Anwendung der quantitativen Methoden in der Linguistik.