Homogeneity and heterogeneity within language(s) and text(s): Theory and practice of word length modeling

Peter Grzybek, Graz

"*Homogeneity in language is merely an idealization, heterogeneity is the 'normal' state and it results from processes all of which are stochastic.*"
Altmann (1987: 231)

0. Introduction

This contribution[1] attempts to shed some light on the consequences the observation of homogeneity and/or heterogeneity of language, or linguistic data, has for the theoretical modeling process in a quantitative linguistics framework. Starting with an introductory discussion of these two key terms (1), the major points of this discussion will then be amplified with regard to concrete linguistic data. To this end, reference will be made to the theory of word length in a synergetic context (2). Given that word length frequencies are regularly organized, it will then be shown that there are systematic variations not only across languages, but also within languages, specific text types, and even within individual texts (3). On the basis of these findings, it will finally be discussed, how such systematic heterogeneities can be taken into account in theoretical modeling, asking for homogeneity (4).

1. Homogeneity and Heterogeneity

The concept and assumption of language being principally characterized by 'homogeneity' has a long history in the development and understanding of linguistics.[2] It may be said to have started with Herder's definition of nation as a community tied together by a common, uniform and, therefore, homogeneous language, and it is also at the basis of Karl Wilhelm von Humboldt's *Über die*

[1] I am sincerely grateful to Gabriel Altmann, Reinhard Köhler, Benedikt Szemerenyi, Bernhard Wälchli, and Rupert Waldenfels, as readers of an earlier version of this text, for their helpful comments.

[2] Unfortunately, the contribution on „Homogenität und Heterogenität der Sprache: Die Entwicklung der Diskussion im 20. Jahrhundert", which had been announced for the third volume of the representative synoptic reference work *History of the Language Sciences*, ed. by Auroux et al. (2006), still appears there with its title, but is missing from the publication and has never been published.

Verschiedenheit des menschlichen Sprachbaus[3] (1836), where he maintains that nation and language completely coincide, and that, despite all individual heterogeneity, only one language prevails throughout a whole nation, eventually diversified into some dialects to a certain degree. It is particularly characteristic for the Saussurian tradition[4], with its dichotomy of speech (*parole*) and language (*langue*), focusing on the alleged homogeneity of *langue* and displacing any kind of heterogeneity to the realm of *parole*.

The assumption of homogeneity is very convenient both for practical purposes (e.g., school grammars) and theoretical objectives (e.g., grammatical models) in linguistics; it is also advantageous for quantitative analyses of language, homogeneous data being an important pre-condition for many statistical tests. This circumstance is well known as the *ceteris paribus* principle, fundamental to descriptive purposes, theoretical modeling, predictive purposes of scientific inquiry, and the formulation of scientific laws. In the framework of scientific experiments, the *ceteris paribus* assumption is realized by controlling all independent variables other than the one(s) under study, so that the effect of the independent variable(s) under observation on the dependent variable can be isolated. In other words, all other relevant factors are kept constant, and all remaining features – which are regarded to possibly affect the data – are considered to be *external* factors, conceived of as being constant for the sample, at least over the period of observation of the sample.

In reality, however, homogeneous data are but a rare case and difficult to obtain: if at all, they can be drawn only from a single population, concentrating on one or more, but usually not all features of the population. As a consequence, contemporary approaches in the field of linguistics (as in other disciplines, too) have increasingly abandoned previously dominating homogeneous concepts and conceptualizations of language, which had largely excluded variation from the description of linguistic systems for methodological reasons. It is particularly the branch of variational linguistics, basically originating in socio-linguistic approaches, which focuses on the usage and function of particular varieties of language, i.e. not only on sociolects, but also dialects, regiolects, registers, etc. Such variations may result from a whole variety of factors, which are not likely to be reduced to spatial differences; they include group-specific linguistic behavior, situational factors (such as formal vs. informal contexts), stages of language

[3] Humboldt's book was first translated into English under the title of *The Heterogeneity of Language and its Influence on the Intellectual Development of Mankind;* a more recent translation is entitled *On Language. On the Diversity of Human Language Construction and Its Influence on the Mental Development of the Human Species* (Cambridge University Press, 1988, 2nd rev. edition 1999).

[4] Cf. Saussure (1916/59): "Taken as a whole, speech is many-sided and heterogeneous" (ibd., 9); "Whereas speech is heterogeneous, language, as defined, is homogeneous" (ibd., 15), "Taken as a whole, speech cannot be studied, for it is not homogeneous" (ibd., 19).

acquisition, speakers‘ age, language contact, and many others. Quite obviously, the specific object, or ‘justification’, of variational linguistics seems to be primarily motivated, or legitimated, by *extra-linguistic* factors.

As such, variational linguistics might seem to be, at first sight, located at the opposite end of the linguistic spectrum as is linguistic typology, if one sees the latter’s major objective to be the comparative study of languages according to their intrinsic (structural) features. At closer sight, however, linguistic typology, aiming at the description of properties (common to various languages), must take into account the structural diversity of the world’s languages, and thus is concerned with variety, too; in its orientation toward the study of universals, linguistic typology cannot but study (possibly) existing differences between languages and features, on the basis of which languages may then be grouped into classes[5], or attributed to types.[6]

Seen from this point, linguistic typology thus is equally concerned with variation as is variational linguistics – more concretely: with variation across, or between, languages, which is assumed to be not random, but subject to specific regularities, or limitations, linguistic typology in this sense being concerned with the question of how (or to what degree) such limitations allow for meaningful sub-divisions into various linguistic groups and sub-categories.

---

[5] It may be reasonable here to refer to the general distinction between typology and classification. Classification, being based on a set of criteria which may concern each element of the classified set or not, captures all elements of a given set and unambigously attributes each of them to exactly one class. Such classes, as established by classificational typology, have no theoretical implication, since no classification has ever been established by theoretical laws, and they all have been by way of inductive procedures only. In typology, as compared to this – albeit as well being the result of a grouping process – the given elements are attributed to groups (or types) in such a way that the elements within a given type are maximally similar to each other with regard to the relevant features (internal homogeneity), the types at the same time being maximally different (external heterogeneity). Therefore, a typology always aims at some specific question, and it may fulfil, among others, some heuristic function in the process of theory formation, but a typology obtained by way of inductive processes only, can never lead to a theory (cf. Altmann 2008).

[6] If by ‘type’ we understand a group (collection, set, class) of objects sharing specific characteristics (attributes, features, properties), it is obvious that the individual objects (the variants) belonging to one type (the invariant) all must share one or more properties reflected in the type, but that each of the variants may, of course, have additional properties not reflected in the type. In other words, the variants share some properties, but they do not share others; as a consequence, they stand in some kind of homomorphic relation to each other. The (invariant) type, however, contains (reflects) only those features which all variants share. With regard to these features, the type may be considered to be an (abstract) model (i.e., a conceptual construct) of the (concrete) variants, standing in an isomorphic relation to them, in this respect.

From the point of view of quantitative linguistics – which aims at the formulation of (stochastic) laws in the field of linguistics and which, in the closely related and indispensable process of hypothesis formulation, must inevitably refer to theoretical models – both variational linguistics and linguistic typology are equally concerned with variants and invariants, i.e. with classes or types, and variations thereof. As a consequence, they are inevitably concerned with questions of homogeneity and heterogeneity, be that with regard to elements between objects, or within a given object under study.

It is just this circumstance, which in this respect places variational linguistics and linguistic typology into one and the same boat: typology is not possible without variants being attributed to a type, and the assumption of variation makes sense only along the assumption of a common invariant, i.e., a type. In both cases, the question of homogeneity and heterogeneity inevitably comes into play. As such, they can be simply accepted, i.e. they can be taken as given; as soon as the aim is theoretical modeling, however, they necessarily have to be adequately taken into account. It is important to note, in this context, that as soon as variants and variations are studied, this can be done only on the basis of (at least assuming) the existence of some higher-order invariant, i.e. a type, to which they belong: in other words, heterogeneity or homogeneity always refer to some super-ordinate system, not its individual elements.[7]

Homogeneity thus refers to the "sameness" of a set of elements with regard to the property, or properties, of one or more (features of) elements of some super-ordinate system, not to the individual elements (or their features) themselves. In other words, within a given system, elements once being attributed to it, may be regarded to be homogeneous with regard to the system they constitute (eventually along with other, heterogeneous elements); across systems, selected elements may be homogeneous (eventually, again, along with other, heterogeneous elements) with regard to some (abstract, theoretical) super-ordinate system. As a consequence, variants and variations are by definition heterogeneous and can be attributed to a type only on some higher level; different types, in turn, are again heterogeneous, and may eventually conceived of as belonging to some super-type.

We may thus conclude that in any kind of linguistic analysis, we are inevitably concerned with the question of homogeneity and heterogeneity of the linguistic material. The material itself, as our linguistic object, will always be characterized by internal heterogeneity; for the purpose of, and in the process of model building, however, it can and, in fact, must be reduced to (some) relevant aspects, which then allows us to ask the question of homogeneity.

The possible focus on either homogeneity or heterogeneity has been interpreted by Altmann, as early as in the mid 1980s, in terms of levels of analysis;

[7] With regard to linguistics in general, and linguistic typology specifically, this has been very clearly stated by Skalička as early as in 1966, in his seminal essay "Ein typologisches Konstrukt".

it may also, however, *cum grano salis*, be interpreted with regard to the history of linguistics. According to Altmann (1987), homogeneity in language is but an idealization, heterogeneity being the "normal" state, resulting from stochastic processes:

1. The assumption of homogeneity, considering language as a homogeneous whole; it leads to the examination of rules, to determinism, and classification (not going beyond monothetic classes); it uses only nominal (in extreme cases dichotomic) scales; at a more progressive stage of this level one uses for descriptions such methods of qualitative mathematics as algebra, two-valued logic, set theory, the theory of automata, etc.
2. The recognition of heterogeneity, both in synchrony and diachrony, considering language as a diversified whole, leads to the need to somehow order the variation(s) observed.

Recognizing (and accepting) heterogeneity opens the doors in two directions: (a) backwards to homogeneity, using reduction procedures (i.e. norming, boundaries, types, classes, dichotomies, categories, etc.), or (b) forwards to the next level, focusing research on latent mechanisms bringing about the heterogeneity. In the first case[8] (a), we are concerned with some elaborated kind of "homogeneous"

[8] Although absurd at first sight, this tendency seems to be characteristic for corpus linguistics, too, at least for its later developments. Subsequent to its initial emphasis on inductive and empirical methods, concentrating on performance rather than competence, corpus linguistics became increasingly impressed by the notion of 'representativeness', accompanied by the illusion of the 'the-more-the-better principle', which would make it possible to (re-)construe of the (descriptive, or statistical, rather than prescriptive) "norm" of a given language. The naïve assumption was, at least at that time, that a corpus, if only "large enough", is representative of a given language as a whole. There is, however, from a theoretical point of view, a major logical flaw in this argumentation, due to the inappropriateness of the law of large numbers in the field of linguistics: the basic dictum of this law, saying that the relative frequency of a random event approximates its probability by the repetition of events, is restricted to the repetition of equivalent events, only – and no individual text can ever be equivalent to some other text, unless it is reduced to specific aspects focused. But in this case we are already concerned with a model of the text, which may be said to be homogeneous to the given language, or rather, to a given model of that language. We are thus again facing the problem of homogeneity and heterogeneity; it turns out that the problem is equally relevant for any study of sub-systems, or sub-models. Therefore, it also concerns so-called "domain-specific" corpora, which do not claim (any more) to represent language as a whole, but specific (thematic) domains of it. But neither language nor any of its specific (sub-)domains can be seen as the simple sum of all texts (to be) produced; therefore, no (random, balanced, domain-specific, etc.) corpus can reasonably be claimed to be representative for something beyond the material observed. Conclusions to be drawn beyond the object observed necessarily ask for a model: in this case, and only in this case, scientific hypotheses may be formulated; else, we are concerned with no more

linguistics, which attempts to grasp the heterogeneity, the 'chaos', by means of sampling small segments of language or by means of homogenization; in the second case (b), more theoretical branches of linguistics, particulary processual and systems theoretical (synergetic) linguistics, try to investigate the laws of this "chaos" (including the boundary conditions being at work), thus leading to the construction of theories and attempting to yield scientific explanations (in a strict understanding of this term). It is here, with language being understood as a process of self-regulation (or a result of this process), that quantitative linguistics comes into play, with its ambition to formulate the laws controlling this process, including the boundary (or antecedent) conditions, which are responsible for a large part of the variation involved (cf. Altmann 1985, 1987).

It is a major concern of the present study to demonstrate this in detail. The practical implications of the problems theoretically outlined above, and the relevance and need to pay due attention to the homogeneity and heterogeneity factor, shall be illustrated in the following empirical analyses. It shall be seen that heterogeneity is far from being relevant for what usually is being conceived of as a variety of language: not only the system of language as a whole, and not only any of its (sub)-systems, but each individual text is, in fact, principally characterized by internal heterogeneity, what represents a crucial methodological problem for any quantitative analysis of language and text.

By way of an example, the following analyses will concentrate on word length. This is by no means to be understood that word length is, or should be, considered to be a crucial (or even the only) factor for linguistic classification and/or typology. In a way, word length can be considered to be an arbitrarily chosen factor here, which could be replaced (or complemented) by many others. Yet, it is not a completely arbitrarily chosen example since the word, and its length, have been in the center of linguistic attention for a long time.

2. Word length: The word in a synergetic framework

Although the study of word length has a more than 150-year long history[9] it was only in the mid-1990s that a theory of word length came to be developed. Such a development was possible, of course, due to the fact that at that time, many "local" studies were available which had not only shown that the frequency with which words of a given length occur in texts, or languages, is not arbitrary, but

---

and no less than the observation and description of delimited objects. And about these objects, empirical (but not theoretical) hypotheses use to be formed; they are helpful for scientific progress, but not sufficient. It is just here, where we find the difference between empirical and theoretical sciences, between statistics of language and quantitative linguistics.

[9] For a survey of this history cf. Grzybek (2005).

rule-based, and that word length is no isolated category in a theory of language, but related to other linguistic units and levels.[10]

What was not clear, if there is a universal model with which word length frequencies can generally be theoretically described (and if so, which model), or if language-specific models are needed (and if so, how they are interrelated): Elderton (1949), for example, analyzing passages from various writers, discussed the geometric distribution with regard to word length in English; as compared to this, Čebanov (1947) propagated the (1-shifted) Poisson distribution, referring to the analyses from 127 Indo-European languages; and Fucks, in the mid-1950s, would speak of a "general law of word-formation" (1955a: 88, 1957: 34), or, more exactly, as the "mathematical law of the process of word-formation from syllables for all those languages, which form their words from syllables" (Fucks 1955b: 209).

An important step in the discussion of possibly adequate distribution models for word length frequencies was Grotjahn's (1982) contribution, who argued in favor of the negative binomial distribution which, under certain circumstances, converges against both the geometric and the Poisson distribution; the importance of Grotjahn's contribution has to be seen in the suggestion that, instead of looking for one general model, one should rather try to concentrate on a variety of distributions which are able to represent a valid "law of word formation from syllables" (ibd., 73).

This idea was later taken up by Grotjahn/ Altmann (1993) and elaborated by Wimmer et al. (1994) and Wimmer/Altmann (1996). The assumption brought forth in these papers was that the frequency of a given class (P_x) is determined by its preceding class (P_{x-1}), thus resulting in the proportionality relation $P_x \sim P_{x-1}$. Further assuming that this relation is characterized by a specific proportionality function $f(x)$, one obtains $P_x = f(x)P_{x-1}$. Depending on which concrete function is chosen, different frequency distribution models are being obtained. In the above-mentioned papers, the function $f(x) = ax^{-b}$ – i.e., the Menzerathian function, well-known to have an important function in linguistic self-regulation – was assumed

[10] After all, it is not by chance, that word length has played a crucial role in Greenberg's (1960) approach to language typology of that time. Notwithstanding the intensive discussions, modifications and improvements of his quantitative approach, going on still today – for discussions and further developments of this issue see: Krupa (1965), Krupa/Altmann (1966), Altmann/Lehfeldt (1973), Kempgen/Lehfeldt (2004), Kelih (2011) – his approach shows the importance which has been attached to the word and its characteristics: according to his definition, an index of synthesis (I_S), i.e. an index for the degree of syntheticity, is defined as the ratio of the number of words (f_W) and the number of morphs (f_M) in a given text: $I_S = f_M / f_W$. Following Krupa (1965), or Krupa and Altmann (1966), it is reasonable to change numerator and denominator – otherwise I_S would theoretically tend to infinity – and to interpret the result as an index of analyticity $I_A = f_W / f_M$, the index of syntheticity consequently equaling $I_S = 1 – I_A$. As can be seen, this index is specifically related to word length, originally being based on the average length of a word, measured in the number of morphemes per word.

to be the basic function, leading to the so-called Conway-Maxwell-Poisson distribution. There is no need to go into details here; what is more important is the fact that this approach provided a good starting point for a flexible system of distributions.

Thus, the function $f(x) = ax^{-b}$ was not the only one taken into consideration; rather a whole system of modifications, extensions, and generalizations was described, resulting in a number of different distribution models.[11]

Later[12], this approach was integrated into Wimmer and Altmann's (2005) even more general "Unified Derivation of Some Linguistic Laws". It would lead too far here to discuss this approach in detail; in short, for a discrete variable *X*, this general approach leads to recurrence formula (1) from which, among many others, the above-mentioned distributions can easily be derived:

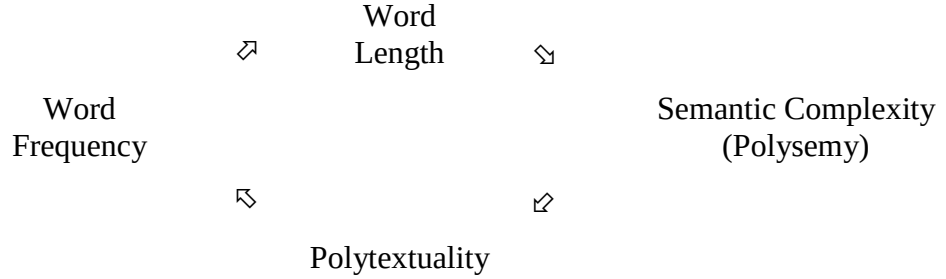
$$(1) \qquad P_x = \left(1 + a_0 + \frac{a_1}{x} + \frac{a_2}{x^2} + \ldots\right) P_{x-1}$$

[11] Thus, to give but a few examples: With $b = 1$ in the basic form mentioned above, one obtains $f(x) = a/x$, leading to the Poisson distribution, with $b = 0$ ($0 < a < 1$), the geometric distribution; from $f(x) = (a+bx)/cx$, the negative binomial distribution results, etc.

[12] At the same time, it had increasingly become clear – not only from structuralist approaches to language, but, first and foremost, from synergetic linguistics – that the word is no isolated entity within a language system. Elaborating on Zipf's works from the 1940s, in which a systematic relation between the frequency and the length of words had already been shown to exist, a number of further relations had been reliably proven to exist, concerning, among others, semantic complexity (polysemy), contextual connectivity (polytextuality), etc.:

1. The more frequent words are, the shorter they tend to be.
2. The shorter words are, the more meanings they tend to have.
3. The more meanings words have, the more likely are they to occur in different (con)texts.
4. In the more different texts a word occurs, the more frequently it tends to be used.

With these selected relations, we are thus facing a circle of interrelations – which, in fact, are much more complex and include many more factors –, being integrated into a synergetic concept (cf. Köhler 1986).



Over the last decades, much empirical evidence has been gathered corroborating hypotheses deduced from this approach. It is not the place here to go into further details; what is important, however, is that this approach can be said to provide the basis of deductive reasoning in quantitative linguistics. As a consequence, it is a matter of boundary conditions, how many and which parameters are needed, and which distribution model results from this. By way of an illustration, Wimmer et al. (1994: 100), referring to individual languages, authorship, and genre, as the three most important factors, have conceived of the situation as a cube (cf. Figure 1).

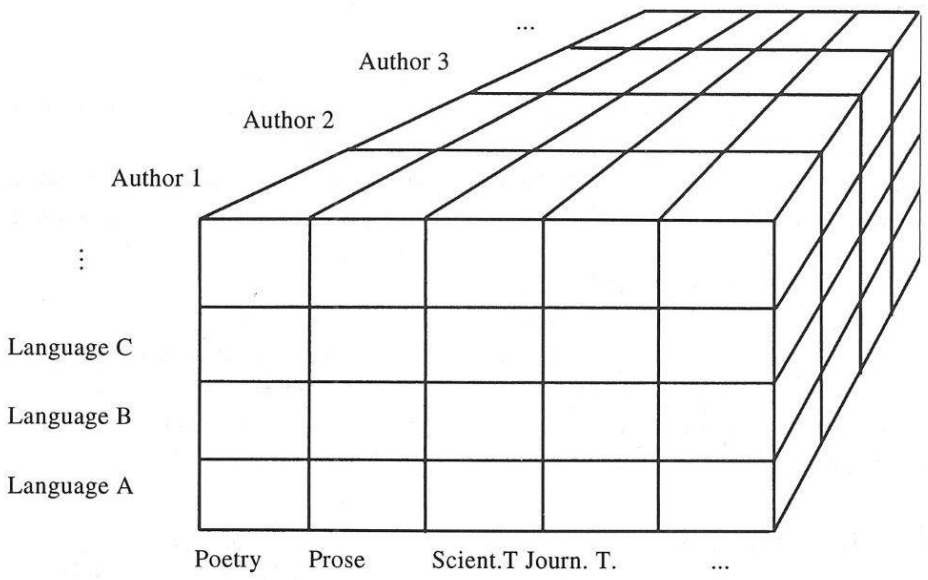


Figure 1: Theory of word length - some boundary conditions (Wimmer et al. 1994)

As can easily be seen, in addition to language, two of the factors, authorship and genre, are of relevance for possible intra-lingual differences, i.e. language-intrinsic word length heterogeneity. This approach has provoked numerous studies on word length, which need not be mentioned here in detail (cf., e.g., Best 1997, 2001). As compared to earlier research, and with regard to the boundary conditions mentioned, these studies are characterized by an interesting tendency: whereas (at least the titles of) part of the papers continue to speak about word length in some language *L* as a whole, others are more specific (or careful) and refer to word length in individual texts of a given author *A* and/or written in some genre *G*.[13]

[13] In this respect, a specific strategy has been to concentrate on the genre of letters, which have been considered to be a prototypical textual representation of a given language, due to its intermediate location between oral and written communication, on

What we are likely to have in the first case, is – given there are intra-lingual differences – some overgeneralized model, based on heterogeneous data. In the second case, we are likely to have more specific models, based on (more) homogeneous data; what we usually do not have, however, are systematic studies of how such specific models (if they differ) or the parameter values of a given model (if there is only one model) relate to each other within a given language. This is, however, a major problem: if there are indeed systematic intra-lingual differences, attention must be paid to them also in inter-lingual comparisons, if one does not want to compare chalk to cheese. Moreover, it is an open question, if and to what degree there are systematic differences not only within a given genre, but also within a given text, which in turn, may well be composed of heterogeneous components.

In the remaining sections, we will first systematically study the problem of language-intrinsic differences, then shifting our attention to text-intrinsic heterogeneities, in order to finally study how such heterogeneities can be dealt with in the process of theoretical modeling.

3. Intra-lingual heterogeneities

3.1. Linguistic data and test material

In order to find out, if systematic word length differences exist within a given language, we need linguistic data as test material which is (or, eventually, can be prepared in such as way that it turns out to be) appropriate for the study of this question. Therefore, the data should not only consist of individual elements (e.g., texts), but also should these elements lend themselves to some kind of systematic grouping (i.e., some text typology).

These groups can either emerge as the result of quantitative analysis – in this case the individual elements are *a posteriori* shown to belong to categories which are heterogeneous with regard to the criterion studied –, or they can represent the starting point, when the individual elements are *a priori* attributed to higher-order classes, or types, which then are tested for systematic differences – in this case, classes which differ with regard to the criterion studied, are considered to be heterogenous, those which do not, as homogeneous. Both approaches are not mutually exclusive – the results may even coincide, although one should not expect this, at least not fully, since we are concerned with one variable only (i.e., word length) here; they simply ask for different methods which will be described and applied further below.

---

the one hand, and the assumption that they represent the outcome of a single, homogeneous (“undisturbed”, “non-interrupted”) process of text generation.

In any case, to pursue these options, we need a corpus[14] of texts and some text typology serving as the higher-order system to which the individual texts can be attributed. From a practical point of view, it is necessary (or at least desirable) to cover the whole textual spectrum of a given language, in order to arrive at systematic results; in that case, at least some implicit knowledge of the textual spectrum and the variety of text types is necessary.

With regard to our objective, it seems reasonable to choose and apply two different text typologies, in order to at least minimially control influences of authoritative decisions in this respect: one of them with maximal abstraction, resulting in minimal specifity and a minimum number of categories (text types), the other one with maximal specifity and, as a consequence, a maximal number of categories. Again, both approaches do not exclude each other, but are to be seen complementary, since the more specified typology should eventually allow for an attribution of its elements to the more general one:

1. As to a maximally specified typology, reference can be made to results from extensive text type research (German: "Textsortenforschung"), where lists with more than 4000 different text types have been provided; these text types[15] are distinguished according to specific communicative-situational functions, which tend to be interpreted in terms of differences in their thematic-propositional or illocutive characteristics (cf., e.g., Adamczik 1995: 255ff.).
2. A minimally specified (and thus maximally reduced) text typology can be seen in the concept of functional styles. This approach originates mainly in Czech functionalism and structuralist positions from the 1930s and 1940s (e.g., Havránek, and others), and it has later been elaborated by Russian scholars (as, e.g., Vinogradov, and others), too.[16] Generally speaking, the concept of functional styles is characterized by the attempt to relate specific stylistic features to extra-linguistic pragmatic or social functions, assuming that specific purposes of language usage influences

[14] It should be emphasized here that within this "corpus" all texts keep their individuality and are not merged into one corpus in the usual understanding of this term. As Orlov (1982) has pointed out some decades ago, any kind of textual combination is, from a theoretical point of view of quantitative linguistics, not an increasingly better approximation to some abstract norm, but a mixture of heterogeneous components – a "pseudo-text", in other words. In this context, the qualitative attribution of individual texts to text types is but tentative, as is the combination of more than one text in some kind of sub-corpus, as long the homogeneity of these texts is not tested and, eventually, proven by adequate statistical methods.

[15] The term 'text type' may be used differently in other contexts; here, it serves as a translation of the term 'Textsorte' as it has become commonly used in German scholarly discourse.

[16] For an informative survey on functional stylistics, including Russian research, see Ohnheiser (1999).

linguistic form.[17] Functional styles have been successfully submitted to quantitative and probabilistic approaches[18], e.g., by Doležel (1964), or Mistrík (1973), who used the "traditional" schema with five major categories of discourse: everyday (colloquial), scholarly (scientific), administrative, journalistic and artistic, the latter inturn being subdivided into literary prose, poetry, and dramatic language (cf. Figure 2).

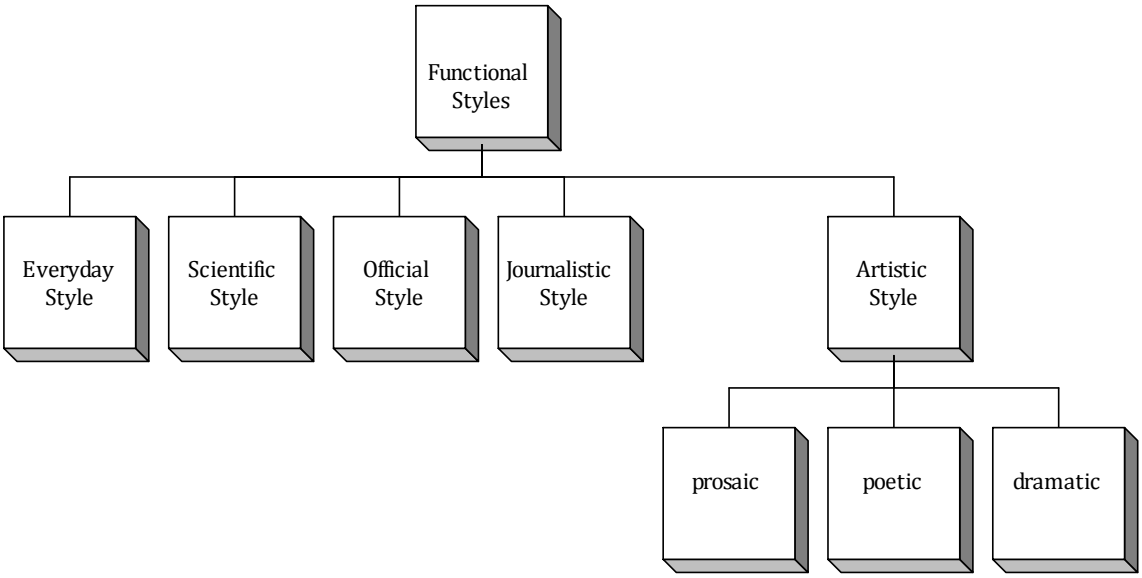


Figure 2: Functional styles (Mistrík 1973: 23ff.)

Although more specific approaches are favored in contemporary text typology (cf. Blühdorn 1990: 218), functional styles continue to play an important role still today, as e.g., in the recently published statistical analyses of the representative Czech National Corpus (Bartoň et al. 2009).[19] For our purposes, it thus seems reasonable and sufficient to refer to these two concepts of text typology. More-

[17] In this respect, the concept of functional style has very much anticipated of what has more recently been discussed in contemporary approaches favoring the notion of 'register' (cf., e.g., Biber 1988, 1995).

[18] In this respect, modern approaches like the ones mentioned before are much more elaborate, both with regard to the number of linguistic variables taken into consideration, and to the amount of quantitative methods applied. For Biber, for example, 'text types' are quantitatively determined on the basis of linguistic similarities, and as the result of extensive statistical analyses; these analyses remain, however, on a descriptive level only, and do not touch upon the question of modeling with regard to a theory, as this is relevant and considered to be crucial in a quantitative linguistics framework.

[19] Detailed results are offered, among others, for word length, separately calculated for three functional styles: scientific, journalistic, and literary prose; word length is counted both in the number of syllables and the number of phonemes per word; additionally, data are given separately for both type and token occurrences.

over, with regard to the compilation of our text data base for the analyses, it seems plausible to choose such text types from the vast amount available, that each category of the less specified text typology (i.e., of the functional styles) is "filled", each by at least ca. 30 texts from at least one text type.[20]

For the sake of illustration, we will use material from a Slovenian text data base described and analyzed in detail elsewhere – cf. Grzybek/Kelih (2005a,b), Grzybek et al. (2006), Kelih et al. (2005): 398 texts from seven different text types were tentatively attributed to functional styles, with four of the seven functional styles being represented by either more than one text type or texts from more than one author, thus representing allegedly homogeneous subgroups of ca. 30 texts each.

It should be emphasized once more that the 398 texts were not merged into a corpus. With regard to the above-mentioned assumption that the genre of letters is a prototypical textual representation of a given language, particular emphasis was laid on different kinds of letters in compiling the text data base. These letters were tentatively attributed to different kinds of text types and, by way of that, to different functional styles: private letters as one instance of everyday communication, open letters as instances of administrative/public and letters to the editor of journalistic communication, chapters from epistolary novels as belonging to literary prose. Additionally, to complement the schema of functional styles, journalistic comments, poems, dramatic texts, and short novels were included, the sum of texts thus summing up to a corpus of 398 items. These texts were not, however, fused into one large corpus – rather, each text was treated in its individuality, thus allowing for tests of the classificatory principles according to the two methods described above. Table 1 represents the text data base in detail.

Table 1
Text basis of 398 Slovene texts

Functional style	**Authors**	**Text types**	**N**
Colloquial	Cankar, Jurčič	Private letters	61

[20] This procedure seems reasonable because the authoritative attribution of text types to functional styles involves the possible methodological problem that it is based on some kind of *a priori* decision only. As such, we are concerned with a qualitative decision, which may well bias the overall result. In fact, as has been shown elsewhere in more detail, such text type attributions to functional styles may be highly subject-dependent: in a study involving 24 experts in text typology, there was high agreement as to some text types, but large disagreement as to others, when subjects attributed specific text types to either more than one functional style at a time, or to different functional styles – for details, see Grzybek/Kelih (2005a,b).

Administration	various	Open letters	29
Journalistic	various	Letters to the editor, Comments	65
Prose	Cankar	Chapters from long stories (‘povest’)	68
	Švigelj-Mérat / Kolšek	Letters from an epistolary novel	93
Poetic	Gregorčič	Versified poems	40
Dramatic	Jančar	Single acts from dramas	42
Total			**398**

3.2. Methods

For the purposes outlined, various statistical methods may be applied. Generally speaking, there are three commonly used approaches, which may be termed quantitative (i), or quantitative-qualitative (ii), respectively. Specifically, we are concerned with:

i. *Clustering methods*, which introduce no qualitative information into the process of classification; rather, they represent some kind of *tabula rasa* principle, introducing specific quantitative information only (such as, in our example, mean word length of a given sample), and aiming at the distinction of sub-groups which in the end will have to be interpreted qualitatively;
ii. *Post hoc* and *discrimination methods*, which are to be understood as specific combinations of *a priori* and *a posteriori* (qualitative and quantitative) principles, which are both based on tentative attributions of the individual samples to groups:
    a. in post hoc analyses, more often than not (but not necessarily) based on the means of the observations, the major question is if specific homogeneous subgroups can be detected among the groups tentatively distinguished *a priori*,
    b. in discriminant methods, the adequacy of tentative *a priori* attributions is tested by first mathematically transforming the variables in order to arrive at a maximal distinction of occurrences, and then calculating the percentage of “correct” attributions – the higher the percentage of “correct” attributions, the better the discrimination is interpreted to be.

Strictly speaking, we need no text typology, if we want to apply method (i) only. Explicit recourse to some text typology is necessary, however, for methods (iia) and (iib); it goes without saying that, in this case, the results to be obtained may at least partly depend on the concrete typology chosen.

3.2.1. Cluster analysis

In a first approach, clustering methods were applied, where no qualitative information is introduced. A usual procedure to determine the optimal number of clusters is the so-called elbow technique, which is based on the mean of the squared errors of an analysis of variance for a give number of clusters (which can be stepwise varied). Table 2 contains the values obtained for 3-8 clusters, which are graphically represented in Figure 3: in the two-dimensional graphic, the number of clusters and the sum of the squared errors are depicted on the *x* and *y* scales; the "best" number of clusters can be seen from that point of the curve, where a salient descent (the 'ellbow') can be observed.

As compared to this, two-step analyses represent an explorative procedure to identify groups within a given data set, various distance measures being used to calculate the (dis)similarities between clusters.

Table 2
Mean of squared errors for different numbers of clusters

Number of clusters	**Mean of squared errors**
2	0.017
3	0.009
4	0.006
5	0.004
6	0.003
7	0.003
8	0.002

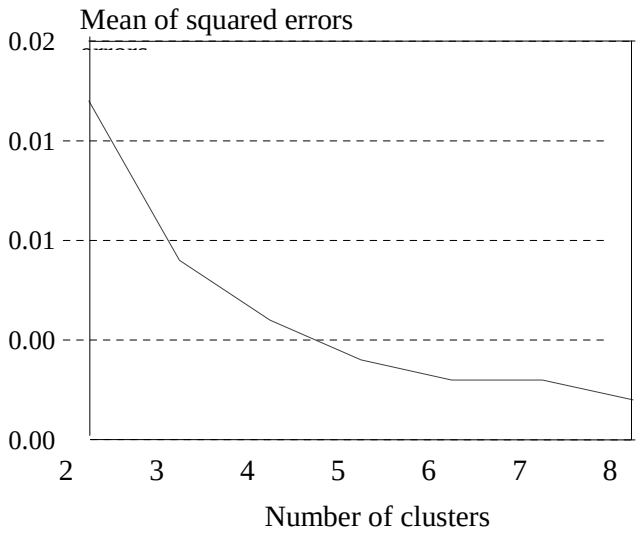


Figure 3: Optimal number of clusters (ellbow technique)

Table 3 shows the results, based on log-likelihood distances.

Table 3
Two-step cluster analyses

Centroids			
		$\bar{x}$	$\bar{s}$
Clusters	**1**	2.4020	0.1293
	2	1.8114	0.0885
	3	2.0450	0.0794
	Combined	1.9889	0.2379

As can be seen, the same result is obtained by both procedures, the visual elbow technique as well as two-step cluster analyses based on log-likelihood distances: accordingly, the "optimal" number of clusters to be distinguished for the material under study turns out to be three. This result is most surprising, since the number of three categories corresponds neither to the number of text types studied, nor to the number of functional styles. In other words, there do seem to be specific textual sub-categories – what is a clear indication of systematic intra-lingual heterogeneity –, but not in agreement with either of the text typologies applied.

3.2.2. Post hoc analysis

Approaching the problem from a different side, post hoc comparisons of means can be run, based on a priori attributions to text types, on the one hand, and quantitative information (in our case: word length averages per text) on the other, in order to identify possibly existing homogeneous subgroups (i.e., without significant differences within the groups, but with significant distinctions between the groups. Table 4 represents the result of these analyses.

Table 4
Post hoc comparisons of means (8 text types, 398 texts)

		Homogeneous subgroups ($\alpha = 0.05$)				
Text Type	*N*	1	2	3	4	5
Poems	40	1.7127				
Short stories	68		1.8258			
Private letters	61		1.8798			
Drama	42		1.8977			
Epistolary Novel	93			2.0026		
Letters to the Editor	30				2.2622	
Comments	35				2.2883	
Open Letters	29					2.4268
Significance		≈1.00	0.37	≈1.00	0.99	≈1.00

As can be seen from Table 4, five homogeneous sub-groups do indeed exist, to which the texts from the eight text types chosen can be attributed. At closer sight, however, some more specific observations raise interpretative problems:

1. With five sub-groups, the number of identified homogeneous sub-groups is different from the number of clusters, obtained in cluster analyses.
2. There is no consistent attribution of text types to functional styles as predicted.
3. The four different letter types fall into four different categories.

In sum, we seem to have homogeneous sub-groups, but neither of the two qualitative typologies applied corresponds to these five subgroups.

3.2.3. Discriminant analysis

In discriminant analyses, individual cases (here: texts) are first attributed to groups (here: text types, or functional styles, respectively) on the basis of specific predictor variables (here: average word length and statistical characteristics derived therefrom[21]), these variables then being submitted to linear transformations, in order to arrive at an optimal discrimination of the cases. However, running discriminant analyses with text types, thus testing the hypothesis "Word length is a variable, which is characteristic of text types", we arrive at the poor result of only 56.3% correct attributions of the texts – what causes us to reject this hypothesis.

A better – though still far from satisfying – result is obtained for discriminant analyses on the basis of functional styles: in this case, in contrast to the assumption of homogeneity of word length within functional styles, we arrive at a still overall poor percentage of 73% correct discriminations.

There are various possible explanations at hand for the overall poor results of the discriminant analyses, e.g.:

a. the tentative *a priori* attributions of the individual texts to text types and/or the attribution of text types to functional styles have been wrong or inconsistent;
b. none of the typologies (i.e., neither the text types distinguished nor the functional styles) provides an adequate (basis for) text typology;
c. there are no consistent subgroups to be distinguished on the basis of word length, which thus turns out to be no good indicator for the demonstration of systematic intra-lingual heterogeneity: the property of word length, used as a basis of classification, is not adequate for the given purpose.

[21] Statistical characteristics derived from the word length frequency distribution, are measures such as variance, dispersion coefficient, skewness, kurtosis, etc., in addition to the mean.

By stepwise (progressive) re-grouping, it can be shown, however, that indeed three general categories can be distinguished, i.e., a number of categories which corresponds to the initial result of our cluster analyses. According to the results, we are concerned with *three discourse types* (as they shall be termed here), which can be distinguished rather clearly on the basis of word length: juxtaposing (i) poetic texts vs. (ii) public (written) speech vs. (iii) private (or oral speech[22]), the outcome is a remarkable percentage of 92.7% correct discriminations. Table 5 represents the results.

Table 5
Three discourse types as a results of discriminant analyses

	Predicted group			Total
	Oral /Private	Written /Public	Verse	
Oral / Private	260	3	1	264
Written / Public	19	75	0	94
Verse	6	0	40	40

Figure 4 offers a graphical illustration of these results.

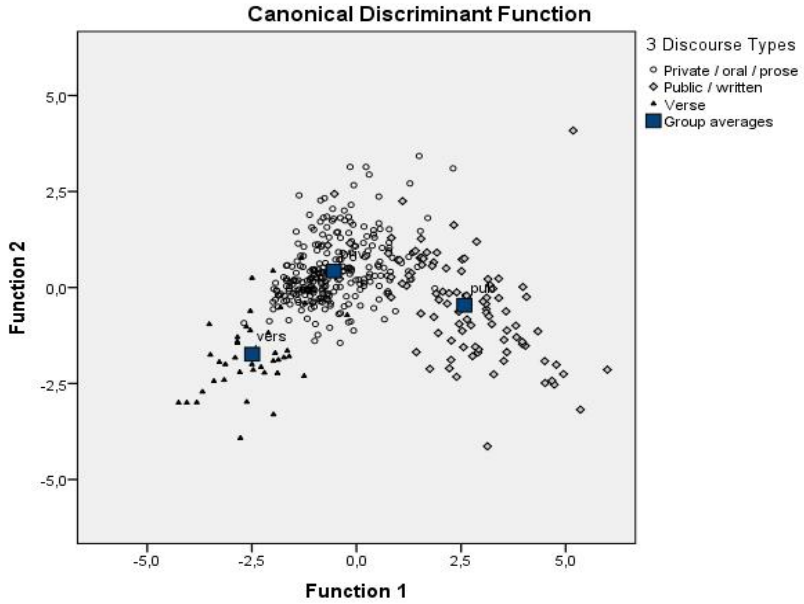


Figure 4: Discrimination of three discourse types

[22] It should be mentioned that the 19the century literary stories analyzed here not only include many dialogues (i.e., fictitious oral speech), but that the whole *ductus* of these texts aims at the illusion of the narrator using oral speech, a phenomenon known as "skaz" in literary theory; both factors might explain why these texts might rather be classified as oral speech (what by no means must be characteristic of literary prose in general).

Summarizing the major results of this section, we can thus conclude that language turns out to be no homogeneous whole, at least not with regard to word length.[23] Rather, there is a large portion of systematic heterogeneity immanent to a given language[24], beyond (or rather below) extra-linguistically motivated categories.[25] Functional styles, albeit the most "radical" kind of intra-lingual text typology, seem to reflect this heterogeneity inadequately; rather, there seem to exist a limited number of more general discourse types which can be distinguished on the basis of word length, obviously even more "trivial" than the maximally reduced functional styles. This, in turn, might be a hint at the conclusion that, due to the "triviality" of these categories, concrete texts from either one of the functional styles or one of the text types, might be composed of mixtures of these categories, what will have to be tested further below, when intra-textual heterogeneity will be at stake.

With this in mind, let us now turn to the questions of variation within a given text type and within one and the same text, again with regard to word length.

3.4. Heterogeneity within text type and texts

As to the study of variation within a given text type, let us analyze, by way of an example, two Russian texts: both belong to literary prose, both are written by one and the same author, Aleksej N. Tolstoj (1882-1945), and both were published within a time span of six years and thus approximately at one and the same period of the author's life. One text, *Гадюка* [The Adder], is a story from 1928, the other one is *Золотой ключик* [*The Little Golden Key*], a story for children from 1936.

Starting with a comparison of average word (x) length for these two texts[26], it turns out that words in the children's text are shorter on the average, with $\overline{x} = 2.31$ ($s = 1.10$), than those in the adults' text, with $\overline{x} = 2.42$ ($s = 1.21$).

[23] Quite similar results have been obtained with studies based on sentence length, or on a combination of word and sentence length jointly (cf. Kelih et al. 2006).

[24] Very similar results have also been obtained with analogical studies on Russian (Friedl 2006).

[25] The observed heterogeneity cannot be explained, by the way, by individual authors' style: in a detailed study on authorship, letters and poems by three Russian poets (A.S. Puškin, A. Achmatova, D. Charms) were analyzed (ca. 30 texts per author and genre, summing up to a total of 190 texts); as a result, it turned out that, with authorship as the discriminant variable, there was a percentage of only 38.4%, as compared to 89.5%, with genre as the discriminating variable (cf. Kelih et al. 2005).

[26] Merging both texts into one "corpus" results in a mean word length of $\overline{x} = 2.35$ ($s = 1.15$). As compared to the idea that corpus construction is an appropriate procedure to "smoothen" heterogeneities and to illuminate a language's "norm", it can easily be seen that actually, such kind of "norm" is but an artificial construct, the corpus in fact turning out to be but a mixed pseudo-text in Orlov's sense (see above).

Since word length frequencies are known not to be normally distributed (see above), a Mann-Whitney *U*-test is in order to test the differences for significance. As the result shows, the differences are highly significant (z = -5.23, $p < 0.001$). In other words: mean word length clearly differs for these two texts, written by one and the same author, belonging to one and the same text type, literary prose.[27]

Given this finding, we can go one step further, showing that heterogeneity may characterize not only relations between two texts of one and the same text type, but also characterizing specific textual subgroups within these texts. To demonstrate this, let us separately analyze the narrative and the dialogical passages of both texts (combined), with regard to average word length, and compare the results for differences between both sub-groups.

Calculating average word length yields in an interesting – though, at second sight, not really astonishing – result: the (combined) narrative passages of both texts are characterized by clearly shorter word length (with $\overline{x} = 2.41$, $s = 1.16$) as compared to the (combined) dialogical passages ($\overline{x} = 2.15$, $s = 1.11$), the difference being highly significant ($z = –16.60$, $p < 0.001$).

Given this observation, it is obvious that average word length of a given text is heavily biased by specifics of text composition, and it is likely to be influenced by the proportion of narrative and dialogical passages contained. Taking this finding into account, it seems reasonable to pay attention to the percentages of these two constituting elements in the children's and the adults' texts, and to compare differences in proportions: of the overall 10590 words in the adults' text, 8120 are represented by narrative[28] passages, and 1527 by dialogues; this corresponds to 76.68% narrative passages and 14.42% dialogues. As compared to this, the children's text contains 10085 words from narrative, and 5291 from dialogical passages, from a total sum of 17470 words; in percentages, this corresponds to only 57.73% of narrative passages and 30.29% dialogues. As a chi square test shows, these differences are highly significant ($X^2 = 1032.55$, $p < 0.001$). Thus, the quantity distribution of narrative and dialogical passages – which, as has been shown above, clearly differ with regard to word length – turns out to heavily influence the overall result.

Yet, the observation of intrinsic heterogeneity is not at its end here. Comparing average word length in narrative passages and in dialogues of the adults' and the children's text, separately for each of the two texts individually, it turns out that these again clearly differ across texts: the narrative passages in the adults' text are significantly ($z = –2.98$, $p < 0.005$) longer ($\overline{x} = 2.46$) than those in

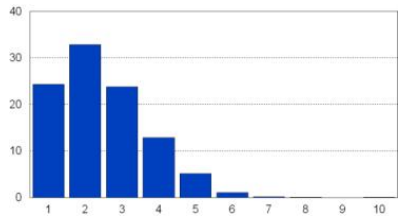
[27] It is a well-known statistical fact that differences in large samples generally tend to be more likely to be significant; however, in case of the non-parametric *U*-test, sample size plays no crucial role, thus indicating the results to be reliable.

[28] No distinction is made here between narrative and descriptive passages; moreover, auctorial narrative sequences preceding (i.e., introducing) or following figures' direct speech, are ignored here.

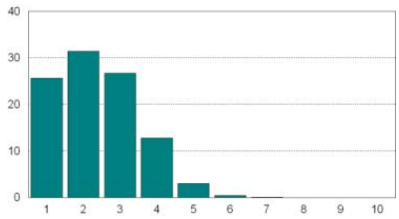
the children's text ($\bar{x}$ = 2.37); furthermore, word length in the dialogical sequences of the adults' text ($\bar{x}$ = 2.21) is significantly shorter than in the children's text ($\bar{x}$ = 2.13), the difference in this case not being significant, however (z = –1.21, p = 0.22).

The observed differences are, of course, a result of differences in frequency distribution – after all, the mean is but one central measure of central tendency. Figures 4a-d illustrate the observed word length frequencies for the four subsamples.

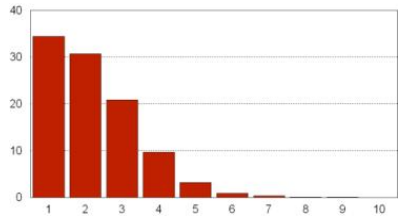
As can easily be seen, the distribution profiles for the narrative and for the dialogical sequences clearly differ, word length for the narrative passages having a peak at two-syllable words, whereas monotonously decreasing for the dialogues.



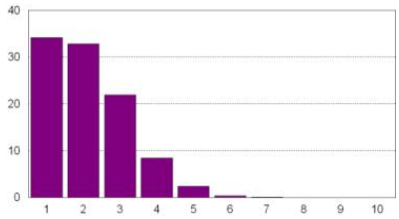
(a) Narrative adults



(b) Narrative children



(c) Dialogues adults



(d) Dialogue children

Figure 4: Empirical word length frequencies in four sub-samples

In any case, despite the seemingly similar profiles of the two dialogical as well as the two narrative sequences across the children's and the adult's texts, differences in average word length are significant, as has been shown above. This is confirmed by the non-parametric Kruskal-Wallis *H*-test for between-group differences with all four groups, which in our case, with the variable 'word length' not following a normal distribution, has to be used for the analysis of variance (ANOVA). As a result, the differences between the four sub-groups are highly significant (X^2 = 285.78, d.f. = 3, p < 0.001). This test result can only indicate the existence of differences, but it cannot identify which of the groups is (or are) responsible for the differences. Therefore additionally computing post hoc tests with all four samples, in order to identify possibly existing homogeneous subgroups, yields the insight that there are no homogeneous subgroups; rather they all fall into a separate category of their own, the three most common post hoc tests (Student-Newman-Keuls, Tukey-B and Scheffé) all equally yielding

high significance, with $p \approx 1.00$. Table 6 represents the results of the post hoc tests.

Table 6
Post hoc tests with mean word length

		Subset for α = 0.05			
Group	*N*	**1**	**2**	**3**	**4**
Dialogue (children)	5291	2.13			
Dialogue (adults)	1527		2.21		
Narrative (children)	10085			2.37	
Narrative (adults)	8120				1.46
Significance		≈1.00	≈1.00	≈1.00	≈1.00

As a result, it thus turns out that not only the two texts are heterogeneous with regard to word length, but that, in addition to this, each of these texts is heterogeneous in itself. Eventually, even more specific sub-samples might be thought of, as e.g., differences between individual speakers, between neutral utterances, questions, and imperatives, and so on and so forth… But instead of further complicating matters, let us draw some preliminary conclusions from the foregoing observations.

3.5. In-Between-Conclusion

Summarizing the results obtained from cluster, post hoc, and discriminant analyses, we may thus say that the principle of heterogeneity goes indeed much farther as is often assumed or taken into account. It seems, any linguistic attempt at describing general "norms" of a language, must ask itself in how far this fact is relevant for the given question and eventually pay due attention to it.

At first sight, the insights gained may seem to be most important for corpus linguistics, particularly when the latter is concerned with theoretical generalizations of empirical results obtained. With corpus linguistics abandoning its "the-more-the-better-principle" – and, by way of that, changing its orientation from establishing the norm of a given language to that of specific domains of it – only a first step seems to be done. Ultimately, any linguistic attempt at (re)-constructing generally valid norms must take into account a major conclusion to be drawn from the observations above, namely that such norms seem to vanish, the deeper one goes into details. In trying to provide homogeneous material – as was said earlier in this text, a necessary pre-condition for statistical testing and reasoning –, the ice gets getting increasingly thinner under the linguists' feet: after the illusion of finding a norm of language as a whole, attention was directed to corpora considered to be "domain-specific", or of "context-related relevance", and it seems attention may, or must further be turned towards "balanced" corpora of specific text types, eventually restricted to specific individual authors, maybe

even from a clearly defined period of time, and so on, and so forth… In the end, we have nothing but the text itself; but even a text is heterogeneous in itself, as could be seen above.

This phenomenon is far from being specific for linguistic objects, and well-known to scientist from many other fields. It seems that under these circumstances, no generalization beyond the object observed is possible any longer. Ultimately, this disillusioning result would make it impossible to do scientific research. The idea to base linguistic research on allegedly prototypical texts, may seem to be a way out; but as could be seen, a single prototype, does not exist, and it has to be chosen, or rather defined, anew, with any question to be pursued. And, what is even more important, if we do not want to restrict ourselves to authoritative qualitative decisions, we tend to know only post hoc, what an adequate prototype is for problem under study.

This raises the final question, how one can deal with these problems in the "everyday practice" of quantitative linguistics, for which the establishment of theoretical models is a sine qua non condition in its research paradigm.

4. Modeling heterogeneities

As has been emphasized above, linguistic objects tend to be principally characterized by heterogeneity, being essential to any kind of linguistic material under study. Yet, with regard to an abstract model, adequate to theoretically describe and eventually explain these data, it is just homogeneity which is needed: data homogeneity is necessary as soon as forming and testing a hypothesis is at stake, which refers to a mechanism one assumes to exist „behind" or "beyond" the data observed.[29]

Data acquisition, in a quantitative linguistics framework, has to be functionally seen as the foundation of theoretical conclusions, with the aim to develop stochastic laws, and quantification is but a necessary step in the logical sequence of scientific steps (cf. Altmann 1993) which generally comprise:

1. Qualitative formulation of a hypothesis, which relates to language(s) or text(s), are of empirical relevance and testable;
2. Statistical formulation („translation") of the hypothesis;
3. Empirical testing (retaining / rejecting) the hypothesis;

[29] In this respect, it is important to pay attention to the conceptual distinction of (a) linguostatistics, or statistics of language(s), on the one hand, and (b) quantitative linguistics, or quantitative text analysis, on the other: whereas linguostatistics aims at the description of language(s) and texts, including the number of languages, of speakers, etc., and refers to any statistical description of linguistic phenomena, (b) primarily aims at the formulation of linguistics laws, including theoretical hypotheses to be tested, and thus is, among others, characterized by a different status and function of both data gathering and quantification.

4. Statistical interpretation of the result with regard to the initially formulated hypothesis;
5. Qualitative interpretation.

Quantification thus is not the aim or the outcome of quantitative linguistics, but one necessary step in the course of scientific study.[30] Anyway, the need to base any generalization on homogeneous data, has been explicitly pointed out at the very beginning of this text. Given the theoretical and empirical insights reported above, it turns out, however, that realistically, obtaining (perfectly) homogeneous data is almost impossible (as in other research fields, too).

Generally speaking, in case heterogeneity is observed in (a set) of data, there are two options to deal with it (cf. Altmann 1992):

1. strive for a diversification of the data, aiming at homogeneous subsets, in order to guarantee the *ceteris paribus* condition.
2. integrate heterogeneity into the model, what results, among others, in mixture or composite models;

Let us illustrate this problem, once more using the word length data presented above. Figure 5 represents the data of both texts mentioned above, merged into one "corpus"; since not only narrative and dialogical passages are included, but also auctorial speech accompanying direct speech, the total of words sums up to N = 28060. The second column of Table 7 represents the observed frequencies (f_x) of the individual word length classes (x), graphically represented in Figure 5 by dark grey bars; for the time being, the values in the third column (nP_x) and the light grey bars can be ignored here (see below).

Table 7
Word length frequencies

x	f_x	nP_x
1	7538	7205.73
2	9017	9795.96
3	6982	6658.65
4	3327	3017.41
5	1000	1025.52
6	161	278.83
7	27	63.18
8	6	12.27
9	1	2.09
10	1	0.36

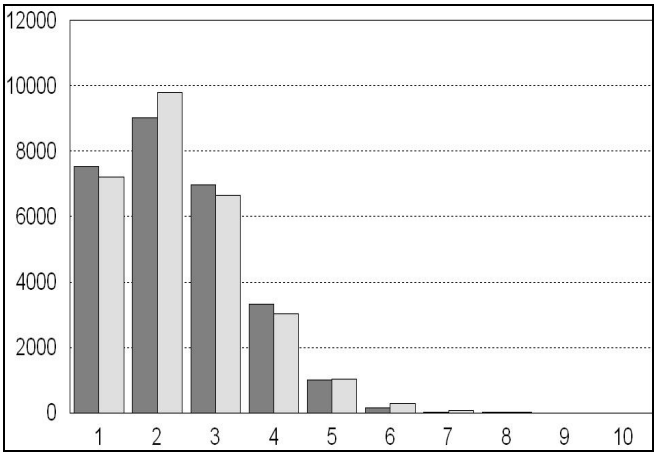


Figure 5: Word length frequencies of two combined texts – observed data and fit of the Poisson distribution (2b)

[30] If Bailey (1991) had been familiar with these principles, and with the theoretical and methodological basics of quantitative linguistics, he could easily give a positive answer to his provocative question "Variation in the data: Can linguistics ever become a science?"

An attempt to find an adequate frequency distribution model for these data may be theoretically based on the general approach discussed above,

$$(1) \qquad P_x = \left(1 + a_0 + \frac{a_1}{x} + \frac{a_2}{x^2} + \ldots\right) P_{x-1}.$$

For $a_0 = -1$, $a_1 > 0$, and $a_i = 0$ for $i = 2,3,\ldots$ we obtain the well-known Poisson distribution (2a):

$$(2a) \qquad P_x = \frac{a^x e^{-a}}{x!} \qquad x = 0,1,2,\ldots$$

Testing the goodness of this model for our word length data, it is reasonable to use this model in its 1-shifted form since, according to our word definition there are no zero-syllable words. We thus arrive at

$$(2b) \qquad P_x = \frac{a^{x-1} e^{-a}}{(x-1!)} \qquad x = 1,2,3,\ldots$$

In fact, with parameter value $a = 1.36$, which is estimated from the data, the fit turns out to yield very good results, as indicated by the discrepancy coefficient[31] $C = X^2/N = 0.0071$. The third column of Table 7 (see above) contains the theoretical values (nP_x) obtained, which are graphically represented by the light grey bars in Figure 5.

Testing the same model for the two texts separately, it turns out, however, that the fit is, in fact, excellent for the adults' text ($C = 0.0057$, with $a = 1.43$), but less good for the children's text ($C = 0.0144$, with $a = 1.33$). Moreover, with regard to the four subgroups (i.e., separately for the narrative and dialogical passages of each text), we see that the model is not only less good for the narrative passages in the children's text, but even has to be rejected for the dialogical passages of the adults' text. The fitting results are represented in detail in Table 8.

[31] The usual goodness of fit test would be the well-known chi square test. Since the X^2 value increases linearly with an increase of sample size, it tends to yield significant result the sooner, the larger the sample is. Since this is the standard case in linguistics, the discrepancy coefficient is preferred, with $C < 0.02$ being interpreted as a good, $C < 0.01$ as a very good fit.

Table 8
Fitting the Poisson distribution to narrative and dialogical sequences

	Adult		Child	
	narrative	dialogue	narrative	dialogue
N	8120	1527	10085	5291
a	1.46	1.20	1.39	1.13
C	0.0039	0.0231	0.0170	0.0070

Thus, with regard to the four subgroups, which we tentatively assume to be homogeneous, and attempting to find a common model for all of them, a usual procedure would include one of the following options – cf. Wimmer/Altmann (1996), Wimmer et al. (1999):

1. To test some 'local' modification.– Usually, in word length studies, it is just the first frequency (f_1), which is modified in one way or another, i.e. for some reason (to be explained), there are "too many" or "not enough" words in this class, thus worsening the overall fit of the model. In such a case, the first probability class (P_1) is modified, i.e. modeled separately, usually being estimated from the observed frequency (f_1). In our case, the Singh-Poisson (4) distribution – which, for α = 1 corresponds to the ordinary (1-displaced) Poisson distribution (2a/b) – might be an adequate model (cf. Wimmer/Altmann 1999: 605f.), in case the assumption above should turn out to be correct.

(4) $$Px = \begin{cases} 1-\alpha+\alpha\cdot e^{-a} & x=1 \\ \dfrac{\alpha\cdot a^{x-1}e^{-a}}{(x-1)!} & x=2,3,4,\ldots \end{cases}$$

2. To test some composite (mixture) model.– Since it cannot be excluded that an allegedly homogeneous subgroup is in fact composed of further heterogeneous components, a mixture of either two different distribution models, or of one and the same with two different weighting factors, might be appropriate. In our case, given the overall adequacy of the Poisson distribution (see above), it seems reasonable to test the Mixed Poisson distribution (cf. Wimmer/Altmann 1999: 417f.) which, for α = 0 or α = 1, again results in the ordinary (1-displaced) Poisson distribution (2a/b):

(5) $$P_x = \frac{\alpha\cdot a^{x-1}e^{-a}}{(x-1)!} + \frac{(1-\alpha)\cdot b^{x-1}e^{-b}}{(x-1)!}, \quad x = 1,2,3,\ldots$$

3. To search for some generalization.– A generalization is a more general model, of which particular sub-models turn out to be special cases, usually with one or more of the general model's parameters being equal to or approximating some limit (0, 1, ∞). In quantitative linguistics in general, and with regard to word length frequency particularly, a well-known generalization of (2a/b) is the (1-displaced) hyper-Poisson distribution (6)

$$(6) \qquad P_x = \frac{a^{x-1}}{{}_1F_1(1;b;a)\,b^{(x-1)}} \quad x = 1,2,3,\ldots$$

which for $b = 1$, results in the ordinary (1-displaced) Poisson distribution (2a/b).

Table 9 summarizes the fitting results for all three data options: (i) the whole corpus, (ii) both texts separately, and (iii) the narrative and dialogical passages in each of the two texts. For all three models, the values of both the parameters and the discrepancy coefficient *C* are given.

As can be seen from Table 9, only for the dialogical passages of the adults' text an improvement can be observed with any of the three modifications, as compared to the ordinary Poisson distribution (cf. Table 7 above). Although the overall results are far from being bad, the relatively worse fit for the children's text, particularly for its narrative passages, is obvious. Interestingly enough, none of the modifications yields crucial improvements for these two sub-samples. This is also reflected in the parameter behavior of the models; for both samples we have parameter $\alpha \to 1$ for the Singh-Poisson distribution, $a = b$ and $\alpha \to 0$ for the Mixed Poisson distribution, and $b \to 1$ for the Hyperpoisson distribution, thus all of them having the ordinary Poisson distribution as special or limiting case, the modifications consequently being of no substantial benefit.

Table 9

Fitting modifications and generalizations of the Poisson distribution

Corpus		**Adults'**	**Children's**	**Narrative**		**Dialogical**	
		(Г.)	**(З. к.)**	**Adults'**	**Children's**	**Adults'**	**Children's**
Singh Poisson							
a	1.40	1.48	1.36	1.50	1.42	1.39	1.21
α	0.97	0.96	0.98	0.98	0.98	0.88	0.94
C	0.0058	0.0031	0.0136	0.0031	0.0166	0.0017	0.0032
Mixed Poisson							
a	1.36	2.27	1.33	1.90	1.39	1.43	1.50
b	1.36	1.42	1.33	1.46	1.39	0.16	1.13
α	0.01	0.01	0.01	0.01	0.01	0.83	0.01
C	0.0071	0.0057	0.0144	0.0043	0.0170	0.0015	0.0070

Hyperpoisson							
a	1.41	1.64	1.32	1.60	1.33	1.88	1.31
b	1.07	1.29	1.00	1.18	0.93	2.01	1.25
C	0.0069	0.0036	0.0144	0.0036	0.0168	0.0021	0.0049

Given these findings, it seems reasonable to tackle the problem differently, starting "from the bottom", i.e., searching for an adequate model covering the sub-samples, first, and only then extending the findings to the complete texts, and to the corpus. Thus, re-analyzing the data, it turns out that a specific modification of the well-known binomial distribution (7)

$$(7) \qquad P_x = \binom{n}{x} p^x q^{n-x} \qquad x = 0,1,...,n; \qquad 0 \le p \le 1,\ q = 1-p$$

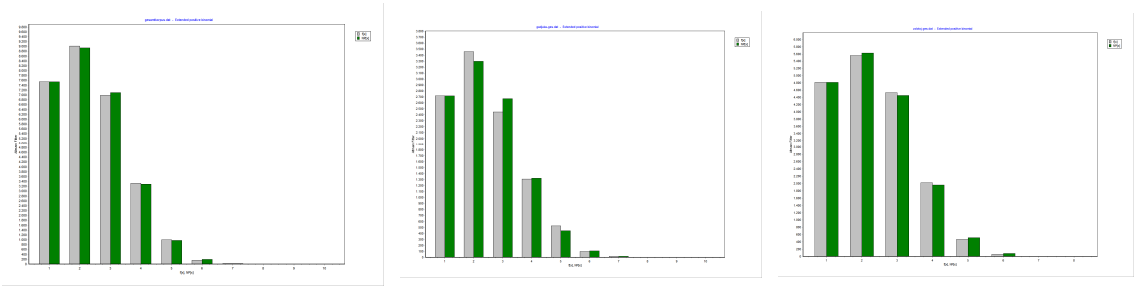
is an excellent model for each of the four narrative and dialogical sub-groups. The binomial distribution (7) can be derived from (1) with $a_0 < -1$ and $i = 2,3,\ldots$ Its modification consists of (a) a left-truncation (which is reasonable, since there are no 0-syllable words, according to our word definition), and (b) a special treatment of the first frequency class P_1 (which would be a hint that it is just the 1-syllable words, which tend to be used in specific ways, asking for some qualitative interpretation). We are thus concerned with the extended positive binomial distribution (cf. Wimmer/Altmann 1999: 148)

$$(8a) \qquad P_x = \begin{cases} 1-\alpha & x = 0 \\ \dfrac{\alpha \binom{n}{x} p^x q^{n-x}}{1-q^n} & x = 1,2,3...,n \end{cases}$$

which in our case is to be used in its 1-displaced form:

$$(8b) \qquad P_x = \begin{cases} 1-\alpha & x = 1 \\ \dfrac{\alpha \binom{n}{x-1} p^{x-1} q^{n-x+1}}{1-q^n} & x = 2,3,4,...n+1 \end{cases}$$

It yields excellent fitting results not only for the four sub-groups, but also for the two texts, and for the complete corpus (with $C < 0.005$ in all cases). This can clearly be seen from the graphical illustrations in Figures 6a-c which show the results for the two text and the combined corpus



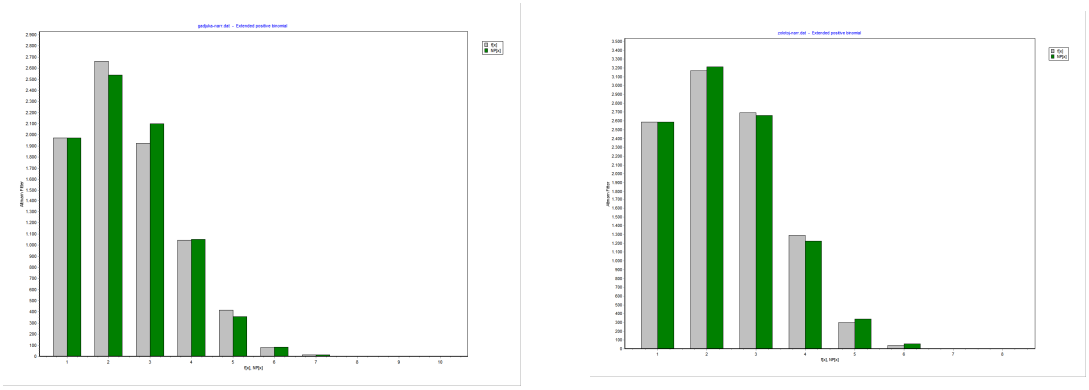
(a) Corpus
$N = 9$, $p = 0.17$, $\alpha = 0.73$
$C = 0.0008$

(b) *Гадюка*
$N = 13$, $p = 0.12$,
$\alpha = 0.74$, $C = 0.0045$

(c) *Золотой ключик*
$N = 7$, $p = 0.21$, $\alpha = 0.72$
$C = 0.0007$

Figure 6: Fitting the extended positive Poisson distribution (texts and corpus)

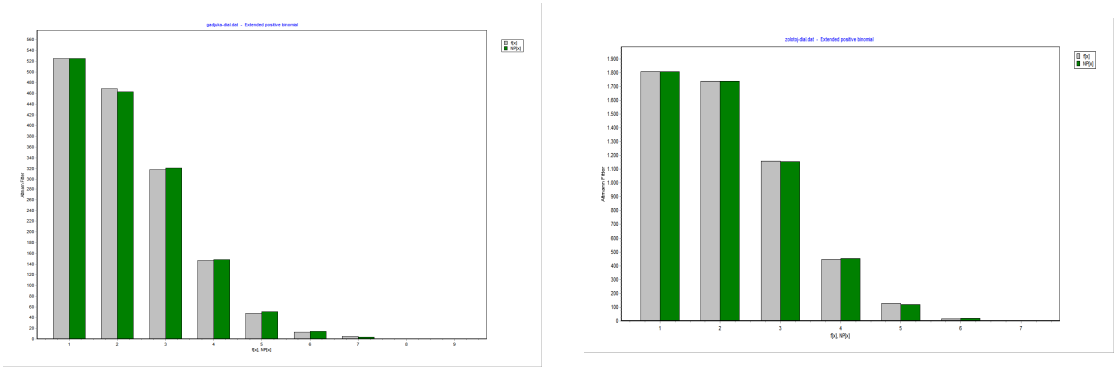
Figures 7a-d present the results for the four narrative and dialogical subgroups. As can be seen, this model is able to grasp all samples equally well, despite the obviously different profile of narrative and dialogical sequences: with $\alpha \approx f_1$ in all cases, and $\alpha \approx 0.75$ for the narrative and $\alpha \approx 0.65$ for the dialogical passages, parameter p of this modified binomial model ranges from $0.12 \leq p \leq 0.21$. Interestingly enough, the model for the dialogical passages in the adults' text slightly deviates from all others, with $n \to \infty$ and $p \to 0$, thus converging to the (extended positive) Poisson distribution.



(a) / (b) Narrative

(b) *Гадюка*
$N = 12$, $p = 0.13$, $\alpha = 0.75$, $C = 0.0043$

(c) *Золотой ключик*
$N = 7$, $p = 0.22$, $\alpha = 0.74$, $C = 0.0014$



(c) / (d) Dialogical

(c) *Гадюка*
$N = 940$, $p = 0.0015$, $\alpha = 0.66$,
$C = 0.0017$

(d) *Золотой ключик*
$N = 10$, $p = 0.13$, $\alpha = 0.66$
$C = 0.0003$

Figure 7: Fitting the extended positive Poisson distribution (four subsamples)

4. Summary and Perspectives

This contribution has started from the assumption that not only are languages different, but that also languages are principally characterized by intrinsic heterogeneity; homogeneity and heterogeneity can only be obtained by way of abstract reduction to specific features under observation, and with reference to some super-ordinate system, or model, concentrating on these features.

Both deductive methods in the Saussure-Chomsky tradition and contemporary approaches favoring inductive methods are doomed to failure in their attempts to arrive at a theory of language, adequately taking into account, among others, variation within language(s), as long as they do not integrate both procedures in an abductive approach, including the formulation of testable hypotheses.

From a quantitative linguistics point of view, linguistic variation is in important object to be studied, which cannot be reduced to extralinguistic factors, but must be understood as the effect of boundary conditions of more general laws, which thus are local specifications, or modifications, of more general language regulations, and which today can already be deduced from a general theoretical concept. Much empirical evidence has been gathered over the last decades, corroborating hypotheses deduced from the “Unified derivation of some linguistic laws”, developed by Wimmer and Altmann (2005, 2006).

By way of an illustrative example, the present contribution demonstrates these principles and procedures with regard to word length, for which systematic varieties are been proven to exist not only within language as a whole, but within specific text types and individual texts, and for which the establishment of the-

oretical frequency distributions are discussed, which attempts to pay due attention to the problems outlined.

References

Adamczik, Kirsten (1995): *Textsorten – Texttypologie. Eine kommentierte Bibliographie.* Münster: Nodus.

Altmann, Gabriel (1985): "On the dynamic approach to language." In: Ballmer, Thomas T. (ed.), *Linguistic Dynamics. Discourses, Procedures, and Evolution*. Berlin: de Gruyter; 181-189.

Altmann, Gabriel (1987): "The levels of linguistic investigation", in: *Theoretical Linguistics, 14; 227-239.*

Altmann, Gabriel (1992): „Das Problem der Datenhomogenität." In: Rieger, Burghard (ed.), *Glottometrika 13*. Bochum: Brockmeyer; 287-298.

Altmann, Gabriel (1993): "Science and linguistics." In: Köhler, Reinhard; Rieger, Burkhart B. (eds.), *Contributions to Quantitative Linguistics*. Dordrecht, NL: Kluwer Academic Publishers; 3-10.

Altmann, Gabriel (2008): „Methodologische Probleme der Sprachtypologie." In: Altmann, Gabriel; Zadorožna, Iryna; Matskulyak, Yuri (eds.), *Проблеми загального, германського та слов'ясьнкого мовознавства до 70-риччя профессора В.В. Левицького. Problems of General, Germanic and Slavic Linguistics. Papers for 70-th anniversary of Professor V. Levic'kij*. Černivci: Books–XXI; 98-105.

Altmann, Gabriel; Lehfeldt, Werner (1973): *Allgemeine Sprachtypologie. Prinzipien und Messverfahren* München: Fink.

Antić, Gordana; Stadlober, Ernst; Grzybek, Peter; Kelih, Emmerich (2006): "Word Length and Frequency Distributions in Different Text Genres." In: Spiliopoulou, Myra; Kruse, Rudolf; Nürnberger, Andreas; Borgelt, Christian; Gaul, Wolfgang (eds.), *From Data and Information Analysis to Knowledge Engineering*. Heidelberg, Berlin: Springer, 310-317.

Auroux, Sylvain; Koerner, Ernst F.K.; Niederehe, Hans-Josef (eds.) (2006): *History of the Language Sciences.* Berlin, New York: de Gruyter. [= Handbücher zur Sprach- und Kommunikationswissenschaft; 18/3]

Bailey, Charles-Jaimes N. (1991): *Variation in the data: Can linguistics ever become a science?* Kea'au, HI: Orchid Land.

Best, Karl-Heinz (ed.) (1997): *Glottometrika 16. The Distribution of Word and Sentence Length*. Trier: wvt.

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

Biber, Douglas (1988): *Variation across speech and writing*. Cambridge etc.: Cambridge University Press.

Biber, Douglas (1995): *Dimensions of register variation: a cross-linguistic comparison*. Cambridge etc.: Cambridge University Press.

Blühdorn, Hardarik (1990): „Korpuslinguistische Befunde als Ausgangspunkt für eine modifizierte Funktionalstilistik – Anregungen zu einer Neuaufnahme der Diskussion", In: *Linguistische Berichte, 127*; 217-231.

Čebanov, Sergej G. (1947): "O podčinenii rečevych ukladov 'indoevropejskoj' gruppy zakonu Puassona" [= On Conformity of Language Structures within the Indo-European Familiy to Poisson's Law], in: *Doklady Akademii Nauk SSSR / Comptes Rendus (Doklady) de l'Adadémie des Sciences de l'URS*, 55/2; 99-102.

Bartoň, Tomaš; Cvrček, Václav; Čermák, František; Jelínek, Tomáš; Petkevič, Vladimír (2009): *Statistyky čestiny*. Praha: *Lidové Noviny*.

Doležel, Lubomír (1964): „Verojatnostnyj podchod k teorii chudožestvennogo stilja", in: *Voprosy jazykoznanija 1*; 19-29.

Elderton, William P. (1949): "A Few Statistics on the Length of English Words", in: *Journal of the Royal Statistical Society, series A (general)*, 112; 436-445.

Friedl, Alexander (2006): *Untersuchungen zur Texttypologie im Russischen*. M.A. Thesis, Graz University.

Fucks, Wilhelm (1955a): *Mathematische Analyse von Sprachelementen, Sprachstil und Sprachen.* Köln/Opladen. [= Arbeitsgemeinschaft für Forschung des Landes Nordrhein-Westfalen; 34a]

Fucks, Wilhelm (1955b): „Theorie der Wortbildung", in: *Mathematisch-Physikalische Semesterberichte zur Pflege des Zusammenhangs von Schule und Universität, 4*; 195-212.

Greenberg, Joseph H. (1960): „A Quantitative Approach to the morphological typology of language", in: *International Journal of American Linguistics*, 26; 178-194.

Grotjahn, Rüdiger (1982): „Ein statistisches Modell für die Verteilung der Wortlänge"', in: *Zeitschrift für Sprachwissenschaft, 1*; 44-75.

Grotjahn, Rüdiger; Altmann, Gabriel (1993): „Modelling the Distribution of Word Length: Some Methodological Problems." In: Köhler, Reinhard; Rieger, Burghard (eds.), *Contributions to Quantitative Linguistics.* Dordrecht, NL: Kluwer Academic Publishers; 141-153.

Grzybek, Peter; Stadlober, Ernst; Kelih, Emmerich; Antić, Gordana (2006): "Quantitative Text Typology: The Impact of Word Length." In: Weihs, Claus; Gaul, Wolfgang (eds.), *Classification. The Ubiquitous Challenge*. Heidelberg, New York: Springer; 53-64.

Grzybek, Peter (2005): "History and Methodology of Word Length Studies. The State of the Art." In: Grzybek, Peter (ed.), *Contributions to the Science of Text and Language. Word Length Studies and Related Issues*. Dordrecht, NL: Springer; 15-90. [Text, Speech and Language Technology; 31]

Grzybek, Peter; Kelih, Emmerich (2005a): „Empirische Textsemiotik und quantitative Text-Typologie." In: Bernard, Jeff; Fikfak, Jurij; Grzybek,

Peter (eds.), *Text & Reality. Text & Wirklichkeit.* Ljubljana, Wien, Graz: ZRC; 95-120.

Grzybek, Peter; Kelih, Emmerich (2005b): „Textforschung: Empirisch!“ In: Banke, Julia K.; Dumont, Björn; Schröter, Anke (eds.), *Die Leipziger Text-Tage.* Leipzig: FSR; 13-34.

Humboldt, Karl Wilhelm von (1836): *Über die Verschiedenheit des menschlichen Sprachbaues und ihren Einfluß auf die geistige Entwickelung des Menschengeschlechts.* Berlin: Königliche Akademie der Wissenschaften.

Kelih, Emmerich (2011): „Zum Analytismus und Synthetismus in den slawischen Sprachen: Morphologische Wortstrukturen in Paralleltexten.“ In: *Polyslav 14*. [In print]

Kelih, Emmerich; Antić, Gordana; Grzybek, Peter; Stadlober, Ernst (2005): “Classification of Author and/or Genre? The Impact of Word Length.” In: Weihs, Claus; Gaul, Wolfgang (eds.), *Classification. The Ubiquitous Challenge.* Heidelberg, New York: Springer; 498-505.

Kelih, Emmerich; Grzybek, Peter; Antić, Gordana; Stadlober, Ernst (2006): “Quantitative Text Typology. The Impact of Sentence Length.” In: Spiliopoulou, Myra; Kruse, Rudolf; Nürnberger, Andreas; Borgelt, Christian; Gaul, Wolfgang (eds.), *From Data and Information Analysis to Knowledge Engineering*. Heidelberg, Berlin: Springer; 382-389.

Kempgen, Sebastian; Lehfeldt, Werner (2004): „Quantitative morphologische Typologie.“ In: Booij, Geert E.; Lehmann, Christian; Mugdan, Joachim (eds.), *Morphologie. Ein internationales Handbuch zur Flexion und Wortbildung.* 2. Halbband. Berlin, New York; 1235-1246.

Köhler, Reinhard (1986): *Zur synergetischen Linguistik: Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.

Köhler, Reinhard (2005): „Properties of lexical units and systems.“ In: Köhler, Reinhard; Altmann, Gabriel; Piotrowski, Rajmund G. (eds.), *Quantitative Linguistik · Quantitative Linguistics. Ein internationales Handbuch · An International Handbook.*. Berlin, New York: de Gruyter; 305-312.

Krupa, Viktor (1965): „On quantification of typology“, in: *Linguistics*, 3/12; 31-36

Krupa, Viktor; Altmann, Gabriel (1966): “Relations between typological indices”, in: *Linguistics* 4/24; 29-37.

Mistrík, Jozef (1973): *Exakte Typologie von Texten.* München: Sagner.

Ohnheiser, Ingeborg (1999): „Funktionale Stilistik.“ In: Jachnow, Helmut (ed.), *Handbuch der sprachwissenschaftlichen Russistik*. Wiesbaden: Harrassowitz; 660-686.

Orlov, Jurij K. (1982): „Linguostatistik: Aufstellung von Sprachnormen oder Analyse des Redeprozesses? (Die Antinomie ‚Sprache–Rede‘ in der statistischen Linguistik)”. In: Orlov, Jurij K; Boroda, Moisej G.; Nadarejšvili, I.Š., *Sprache, Text, Kunst. Quantitative Analysen*. Brockmeyer: **Bochum: Brockmeyer; 1-55.**

Saussure, Ferdinand de (1916): *Course in General Linguistics*. New York: The Philosophical Library, 1959.

Skalička, Vladimír (1966): „Ein «typologisches Konstrukt».“ In: *Travaux linguistiques de Prague,* 2; 157-163.

Wimmer, Gejza; Altmann, Gabriel (1996): “The theory of word length: some results and generalizations. In: Schmidt, Peter (ed.), *Glottometrika 15. Issues in General Linguistic Theory and The Theory of Word Length.* Trier: wvt; 112-133.

Wimmer, Gejza; Altmann, Gabriel (1999): *Thesaurus of univariate discrete probability distrubutions.* Essen: Stamm.

Wimmer, Gejza; Altmann, Gabriel (2005): “Unified derivation of some linguistic laws.” In: Köhler, Reinhard; Altmann, Gabriel; Piotrowski, Rajmund G. (eds.), *Quantitative Linguistik · Quantitative Linguistics. Ein internationales Handbuch · An International Handbook.* Berlin, New York: de Gruyter; 791-807.

Wimmer, Gejza; Altmann, Gabriel (2006): “Towards a Unified Derivation of Some Linguistic Laws.” In: Grzybek, Peter (ed.), *Contributions to the Science of Text and Language: Word Length Studies and Related Issues.* Dordrecht, NL: Springer; 329-337.

Wimmer, Gejza; Köhler, Reinhard; Grotjahn, Rüdiger; Altmann, Gabriel (1994): “Towards a Theory of Word Length Distribution“', in: *Journal of Quantitative Linguistics, 1/1*; 98-106.

Wimmer, Gejza; Witkovský, Viktor; Altmann, Gabriel (1999): “Modification of Probability Distributions Applied to Word Length Research“, in: *Journal of Quantitative Linguistics, 6/3*; 257-268.