Oxford Handbooks Online

Word Length

Peter Grzybek

The Oxford Handbook of the Word *(Forthcoming)*
Edited by John R Taylor

Online Publication Date: Dec 2014

Subject: Linguistics, History of Linguistics
DOI: 10.1093/oxfordhb/9780199641604.013.37

Abstract and Keywords

This chapter concentrates on word length, emphasizing relevant quantitative and synergetic approaches. Alternative units for measuring word length are discussed with regard to their usability, as well as the influence that different kinds of material may have on studying word length. In addition to presenting some basic descriptive statistical characteristics, this contribution shows that word length is a substantial and central phenomenon for a comprehensive theory of language. It is shown, first, that the way in which words of a given length occur in linguistic material is not chaotic, but follows clearly defined, law-like regularities; and second, that word length is not an isolated category within the linguistic system, but is closely interrelated to other properties of the word, as well as of other linguistic units, levels, and structures. Theoretical models are discussed, concerning not only these interrelations, but sequential text analysis and frequency distributions.

Keywords: word length, quantitative linguistics, synergetic linguistics, descriptive statistics, linguistic units, linguistic levels, linguistic structures, sequential text analysis, frequency distribution, theory of language

1 Introduction: Length in Linguistics

Studying the word, as well as other linguistic units, requires quantitative as well as qualitative approaches. Taking language to be a system of rules, or of structures and functions, one might be tempted to assume, erroneously, that we are concerned not with quantities but with qualities, and that we could arrive at a theory of language by way of qualitative methods only; one might also object that language does not (or at least not in all of its aspects) lend itself to quantification. Such objections are, however, but transfers from epistemology to ontology, falsely assuming that qualities and quantities ‘naturally’ exist as such, in and by themselves. In fact, both quantitative and qualitative categories are but abstractions of the mind with which we attempt to grasp the external world; ultimately, we do not quantify external phenomena, but our models thereof (Altmann 1978, 1993).

Length is a quantitative category; it is a property which we can ascribe to a linguistic (or any other physical) object and which can in principle be measured by reference to the dimensions of time and/or space. With regard to word length, it may be useful terminologically and conceptually to distinguish ‘length’ from the closely related concepts ‘duration’ and ‘complexity’, reserving ‘duration’ for the temporal quantification of an event’s unfolding.

Whereas duration is the result of measuring the time course of an event, length and complexity imply spatial measures. Measuring an object’s length involves a spatial perspective along one (horizontal) dimension, length being measured in terms of the number of equivalent components in sequential order which make up the object and serve as its measuring units. In comparison, a linguistic object’s density, or complexity, concerns the number of its elements, their relations to each other, as well as the functions of these relations, not taking account of horizontal extension or sequential order. In practice, measuring the length of a linguistic unit implies the counting of its constituent components, in sequential order, which presupposes the identification of these discrete (linguistic)

units; moreover, all components in this context should be structurally equivalent, i.e. they should belong to one and the same structural level.

With these definitions, the categories of duration, length, and complexity are likely to play different roles in the analysis of spoken vs. written forms of language(s). Given the interaction between written and spoken language, however, complex as they may be for different languages, it becomes obvious that the decision in favour of any one of these concepts is rather a matter of research interest.

Against the background of these introductory remarks, this chapter will be organized as follows. First, some basic definitions are addressed (section 2), concerning the word and the constituent components that serve as measuring units. Next, a number of relevant distinctions are made as to the concrete material serving as a basis of word-length studies; at issue here are distinctions of texts (of different kinds) vs. corpora vs. dictionary material, of word forms vs. lemmata, of types vs. tokens (section 3). Then follow some basic statistical descriptive characteristics of word length (section 4). The major part of the chapter concentrates on theoretical issues, concerning models for word-length frequencies, sequential and positional aspects, and relations of word length to other linguistic phenomena (section 5). Finally, section 6 addresses practical aspects of word length studies, such as the contribution of word length to author identification and to text readability.

2 Definitional Aspects: The Word and its Measuring Units

As is well known since Saussurean times, there are no positive facts in language: the categories applied in linguistics, far from being God-given truths, are the results of authoritative decision or common agreement. From a theoretical perspective, any such definition functions as an axiom, and any result obtained depends on the initial settings. Although the situation is not much different in other disciplines, there is a tendency in linguistics to adhere to specific definitions once they have been made, and to ignore their relative arbitrariness. As a consequence, in word-length studies we should be aware that there is more than one definition of word, as well as of possible measuring units of word length, such as letter, grapheme, syllable, mora, and morpheme.

2.1 What is a Word?

No binding definition of the word, valid for more than one language or even type(s) of language, with possibly different writing systems, can be offered *en passant* in this contribution; for detailed discussions, see Dixon and Aikhenwald (2002) and Wray (this volume). With regard to word length studies, at least for European languages, three operational definitions have been predominantly applied:

a. graphemic/graphematic;
b. phonological (accent group);
c. phonetic-orthographic.

2.1.1

A *graphemic definition* is based on a word's written form, a word being marked by two separators, usually blank spaces or a punctuation mark, occasionally a hyphen. Although this definition is quite practical, particularly for computer-based analyses, a number of problems arise. Irrespective of the fact that there are languages without a written tradition, the definition works only for letter-based (or grapheme-based) scripts, thus excluding languages with other writing systems as well as orthographies without separators, such as Chinese or Thai. Moreover, a graphemic definition is dependent on diachronic developments and changes in orthographic norms (concerning, among other matters, the writing of compounds or the treatment of clitics). Nevertheless, due to its simplicity, a graphemic definition is often favoured by workers in computer and information sciences. In these approaches, it is a matter of stipulation how to deal with hyphenated words (e.g. English: *mother-in-law*) or apostrophied words (English: *that's, isn't, man's*; French: *Jeanne d'Arc*). In any case, it would be consistent with a graphemic definition to measure word length in terms of linguistic units which are realized in written form, such as letters or graphemes. The result will be, of course, an analysis of written language, which, on account of language-specific orthographic rules and specific relations between written and spoken language, may deviate substantially from analyses based on other definitions of the word, the more so since no intermediate constituting levels (such as syllables or morphemes) are taken into consideration. More importantly, a graphemic definition may give contradictory results if

word length is measured in units other than written ones, such as morphemes or syllables—one problem being zero-syllable words, such as the vowel-less prepositions in Slavic languages (e.g. 'к', 'с', 'в' in Russian), which would be counted as words in their own right.

2.1.2

The existence of zero-syllable words and similar problems are avoided in definitions which refer to the phonetic, phonological, or prosodic criteria of spoken language, integrating and emphasizing performance factors of pronunciation. *Phonological word definitions* thus refer to a string of phones or phonemes (or, in related approaches, to their written equivalents), which behave as units for phonetic/phonological processes, particularly the location of (lexical) stress or accent. A basic axiom in these approaches is that a phonological word carries only one (primary) stress. Since English prepositions such as *for* and the definite and indefinite articles *the* and *a(n)* usually are not (though, for pragmatic reasons, may occasionally be) stressed, the sentence *I came for the milk*—consisting of five graphemically defined words—would probably be composed of fewer phonological words, with *I came* and *the milk* each forming one unit, the preposition *for* being attached to the latter expression; also, the treatment of compounds (as separate words or word fusions, with or without hyphenation) presents no major problems in this approach: cf. English *bottle opener* vs. *homeowner* vs. *man-eater*.

2.1.3

Phonetic-orthographic word definitions attempt to combine and balance the technical simplicity of a graphemic approach with linguistic (i.e. phonetic/phonological and morphological) criteria. In this framework, graphemically defined zero-syllable words (see 2.1.1) are interpreted as clitics, with proclitics being merged to the following, and enclitics to the preceding, graphemically defined word. This procedure covers at least some of the orthographic inconsistencies in languages and their writing systems—which, more often than not, are the result of diachronic developments (e.g. Russian *v kratsu* vs. *vkratse*, both variants meaning practically the same: 'in brief', 'briefly').

Not surprisingly, linguistic definitions of the word influence the analysis of word length. A quantitative approach must take account of such influences, and it would seem reasonable to systematically compare the effect of different definitions, which may vary across languages or even within a language, depending e.g., among others, on text type effects. Similar problems are likely to concern measuring units, as will be discussed in the next section.

2.2 Definition of Measuring Units

Word-length measurements will differ depending on which measuring units are chosen; these, in turn, may depend on how the word is defined.

Letters or graphemes might be regarded as adequate measuring units in a graphemic approach. Measuring word length by the number of letters or graphemes is (seemingly) straightforward. The approach is additionally supported by the fact that letters and graphemes are not chaotically distributed, but have their own frequency profile (cf. Grzybek 2007; Grzybek et al. 2009), thus fulfilling a major postulate in quantitative linguistics: that the constituents of a higher-level unit must have their own regular frequency organization.

However, definitional aspects cannot be ignored on this level either. An English word like *shoe* may be considered to consist of four letters; it might also be possible to speak of two graphemes, with the two letters <s+h> counting as one grapheme representing the phoneme [ʃ] and the combination <o+e> representing the phoneme [u:]. Whereas here the components of a grapheme also occur as individual letters in the given alphabet, this need not be the case in other languages. Different definitions are possible in the case of letters containing diacritical characters, such as ä, å, á, ä, ç, õ; these may be considered either as letters in their own right or as combinations of basic characters plus diacritics.

While some of these problems might be solved by measuring word length in terms of phones or phonemes—whether on the basis of a phonetic/phonological transcription or if graphical units are taken to be semiotic representations of spoken language—other problems are likely to arise. After all, it is a matter of linguistic definition what is to be considered a phone, or phoneme; with regard to the analysis of (transcribed) spoken language, additional differences may come into play depending on whether slow, careful, or more casual pronunciation, with

elisions and coalescences, is taken as the norm.

More importantly, neither letters/graphemes nor phones/phonemes are direct constituents of the word: in a traditional structuralist framework, they would be regarded as low-level units, forming syllables or morphemes on the next level, which in turn are then taken to be direct constituents of the word. When linguistics, in contradistinction to information and computer science-based approaches, favours the measurement of word length in terms of the number of direct constituents, there is more than one reason to do so: (a) measurements in terms of indirect constituents are likely to result in a greater amount of variation, thus possibly concealing clear tendencies; (b) measurement fluctuations or inaccuracies, due to definitional aspects, are likely to come into play and to be multiplied the more levels of analysis are at stake; and (c) given that there are control mechanisms which regulate length relations between units of neighbouring levels (see below), the leapfrogging of an intermediate level is likely to obscure (or even disturb) these self-regulating processes.

As a result, measuring word length in terms of the number of syllables or morphemes per word would turn out to be the most appropriate approach, notwithstanding additional problems in defining these units.[1] In any case, although syllables and morphemes measure word length along different scales, there is increasing evidence showing correlations between the results of syllable-based and morpheme-based analyses, at least for the languages studied thus far.

Yet another alternative for measuring word length concerns the number of morae[2] per word. In the context of word-length studies, morae have been used for the analysis of languages like Japanese, not least because here the mora-based approach represents a compromise between the phonetics and the writing system. Given the definitions above, and taking into account that in the context of prosody studies a mora serves a measure for syllable time units, one may rather consider mora-based word-length studies as a mixture of approaches studying duration and length.

3 Material: Sample vs. Text vs. Corpus, Word Form vs. Lemma, Type vs. Token

Once we measure the length, not of a single word but of more than one word, we are able to construct some kind of word-length frequency distribution; this becomes the basis for the derivation of various statistical characteristics (section 4), as well as for the study of theoretical frequency models (section 5). The distribution is likely to vary depending on empirical and methodological factors: on the one hand, the kind of material chosen for analysis will, to one degree or another, influence the results; on the other, the manner of analysis, depending on initial decisions, will modify the outcome.

As to differences concerning the material chosen, one must make a basic distinction between (a) random samples, i.e. randomly chosen parts of texts, (b) complete individual texts, and (c) text combinations, or corpora, composed of different samples and/or texts.

In this context, one must distinguish between the notion of randomness on a linguistic vs. probabilistic understanding of the term. On a linguistic understanding, randomness refers to an arbitrarily chosen text selection. In contrast, a random sample on a statistical (or rather probabilistic) understanding is any selection of a subset of individuals from within a statistical population, made in order to estimate characteristics of the whole population, i.e. to indicate the probability of an item being from the entire population.

An arbitrary text selection can be conceived of as a random sample in statistical terms as well as a complete individual text or some combination of texts or text selections; the crucial question is if the (intended) statistical description of the linguistic material concerns only the material under study, or if conclusions are (to be) made beyond the material observed. The distinction is between descriptive statistics, which confines itself to the material under study, and inferential statistics, which aims at more general statements, based on inferential procedures.

The choice of material is particularly relevant in an inferential framework, where a decision must be made as to what kind of sample material allows for what kind of conclusions. If no conclusions beyond the material under study are intended, choice and control of the material is less relevant and is motivated only by an interest in the given material. As soon as the conclusions to be drawn turn out to be more ambitious and strive to generalize beyond the material under study, attention must be paid to a number of methodological caveats. Any random sample, taken to

be representative of some more encompassing population, denies the existence of intralingual differences, and as soon as such differences are proven to exist, the choice would result in a violation of the assumption of data homogeneity and the ceteris paribus condition. The same holds true for corpus analyses which have long been taken to represent a given language as a whole, since any corpus is but a mixture of heterogeneous texts, or text elements.

In actual practice, it has often been assumed that a given sample, provided that it is 'large enough' (whatever this means in practice and however it may be theoretically based), is characteristic (i.e. 'representative') of a given language as a whole, and can thus serve to establish specific 'norms'; this assumption was particularly prevalent in early corpus linguistics with its 'the more the better' conviction. In less extreme forms, assumptions have been made with regard to some kind of domain-specific, author-specific, or other kind of representativeness, as for example when a sample is considered to be characteristic of individual author styles, text types, chronological periods of a given language, and so on. In this case, the sample-population assumption is related to the assumption of homogeneous sub-groups within the total population.

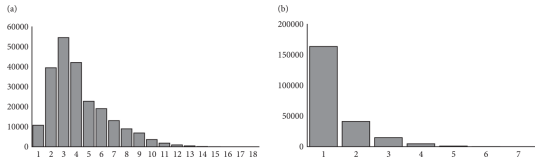
Methodologically speaking, the assumption of data homogeneity is manifested by the desire to control all independent variables other than the one(s) under study, so that the effect of the independent variable(s) under observation can be isolated; in other words, all other relevant factors are (assumed to be kept) constant, and all remaining features, which are regarded as possibly affecting the data, are considered to be external factors, conceived of as being constant for the sample, at least over the period of observation.

In reality, homogeneous data are rare, and this is of specific concern in linguistics. Indeed, a crucial question is whether homogeneity can ever be assumed to exist in language, be that with regard to (a given) language as a whole or to possible (intralingual) subgroups; not only is any combination of texts (a corpus) a fusion of heterogeneous elements (it is no accident that a corpus as a mixture of texts been termed a 'pseudo text': Orlov 1982), each text also differs from any other text, and even within one and the same text the presence of heterogeneous elements is the rule. As a consequence, in order to forestall inadequate generalizations in word-length studies, due attention must be paid to the existence of such intratextual and intralingual heterogeneities. We may note that the genre of letters has long been assumed to be an adequate prototype for a given language's structures (cf. Best 2005), especially since this is a genre on the borderline between spontaneous speech and written language and usually the result of homogeneous acts of text production, less subject to stylistic variation and a posteriori manipulation. Systematic studies have shown, however, that the genre of letters is far from being homogeneous (Grzybek 2013c; Grzybek and Kelih 2006), and different kinds of letter (private letters, open letters, letters to the editor, letters from epistolary novels) are clearly characterized by different word lengths, not necessarily resulting in different theoretical models of word-length distribution (see below).

Not only is the material's quality a crucial factor in the analytical process, so also is its linguistic preparation. Analysing (whatever kind of) text or corpus material necessarily implies the notion of frequency; not all words occur with equal frequency, and given that not all words are of equal length, it is important to decide whether word frequency is taken into account or not. If each lexical appearance is analysed, an individual word's frequency of occurrence plays a major role; if the material is (or has previously been) transformed into word lists, or into dictionaries containing each occurring entity only once, the frequency aspect is deleted. Frequency lists, or frequency dictionaries, represent a special case. A decision on this point can of course not be 'correct' or 'incorrect'; rather, it is a matter of research interest and perspective.

The decision whether or not to take frequency of occurrence into consideration relates to the distinction between types and tokens. In this respect, some important caveats are necessary. First and foremost, it is important to note that the type/token distinction, introduced into scientific discourse by Charles Peirce in the 19th century, is of a rather general kind and concerns not only the lexical level of language, as has often been assumed, but any kind of semiotic entity, lexical items being but one instance. It would therefore be incorrect to identify word forms with tokens and lemmas with types; rather, these are distinctions along two different dimensions. One may be concerned with word form types or word form lemmas, as well as with lemma types and lemma tokens, and decisions on this point will influence word-length measures.

4 Descriptive Characteristics: Object-related



Click to view larger

Fig. 1 . Word-length frequency distribution of *The Portrait of a Lady*.

Within a descriptive approach, statistical measures are derived from a given frequency distribution, in order to characterize it quantitatively. Figs 1a and 1b represent in graphic form the word-length frequency distribution, or 'spectrum', of an English text, the novel *The Portrait of a Lady* by Henry James from 1881. Fig. 1a is based on letter counts per word, Fig. 1b on syllable counts.

It is evident that under both conditions the distribution is not symmetrical, but left-skewed; this is typical for linguistic phenomena in general, not only for word length. It is also clear that the distributions differ in their profiles: not only are there fewer classes under the syllable condition, but the frequencies are monotonously decreasing, whereas there is an initial increase up to a peak of 3-letter words under the letter condition before the remaining frequencies decrease. This aspect is particularly relevant for model-theoretic approaches (see below).

On the basis of empirically observed frequencies, descriptive statistics (or summary statistics) provide specific information about the given distribution in maximally condensed form. They provide measures of location (or central tendency), of dispersion (or variation), and of the shape of the distribution; if more than one variable is analysed, measures of statistical dependence are available. Only the most common characteristics are presented here, using the syllable-based results reported above by way of an example.

Given the absolute frequency f_j of j-syllable words, the total sum of words (N) in a given sample is represented as $\sum_{j=1}^{K} f_j = N$; in our case, N = 225,234 words. On this basis, the relative frequencies h_j of j-syllable words can be computed as $h_j = f_j/N$, the sum of which, ranging from the first element $j = 1$ to the last element $j = K$, equals 1: $\sum_{j=1}^{K} h_j = 1$. Since there are 163,622 one-syllable words, we have $h_1 = 163622/225234 = 0.7261$, which corresponds to 72.61 per cent of all words. On the basis of these frequencies, the arithmetic mean $\bar{x}$, often favoured for characterizing a frequency distribution, can easily be calculated as $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$. In the case of James's novel, we have an average word length of $\bar{x} = 1.3969$. As a minimum of information, any descriptive approach should also give the standard deviation $s = \sqrt{\frac{1}{N} \cdot \sum_{j=1}^{N} (x_i - \bar{x})^2}$ (the square root of the variance s^2) as an essential characteristic of the given sample's measure of variation around the mean value.

Further statistical characteristics can be computed from given frequency data, such as the median, the central moments, the coefficient of variation, the dispersion index, skewness, kurtosis, Ord's criteria, absolute and relative entropy, repeat rate and redundancy. All these measures, in isolation or in specific combinations, may be useful for methods like clustering, discrimination, or post hoc comparisons, when the identification of homogeneous subgroups is at stake.

Compared to such measures, attempts to model word-length frequency distributions as a whole represent a crucial step from descriptive approaches to hypothesis formation and testing, thus building a bridge to theory-oriented approaches.

5 Model-related and Theory-oriented Approaches

The scope of descriptive approaches is to characterize the linguistic material under study as a given product. In comparison, theoretical approaches attempt to provide models which claim relevance not only for the concrete material under study but beyond, and which are thus necessarily based on the formulation of testable hypotheses.

As Altmann (2013: 28), in a synoptic reflection on word-length studies (see also Popescu et al. 2013), has pointed

out, three major conditions must be fulfilled if word length is to contribute to, or be integrated into, a theory of language. These conditions are that

1. word length is not an isolated property;
2. word length underlies language evolution and diversification;
3. word length frequencies are not arbitrary, but abide by laws.

Laws, or law-like regularities, are thus expected to exist with regard to each of these three aspects, with word length as an integral ingredient of a theory of the word within a theory of language.

5.1 From Word-length Spectra to Theoretical Frequency-distribution Models

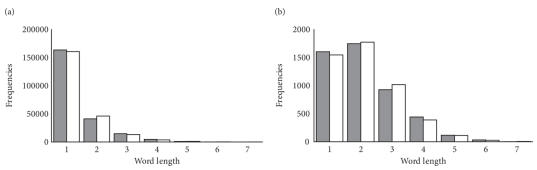
The idea of analysing word-length spectra goes back to the beginnings of word-length studies in the 19th century.[3] After the English logician Augustus de Morgan, in a private letter of 1851 which was published only in 1882, had suggested that questions of authorship might be settled by determining whether one text deals in longer words than another, it was Thomas C. Mendenhall (1887) who initiated systematic word-length studies. As a mathematician, he was familiar with contemporary spectral analysis in physics, and in analogy to this he proposed to go beyond mere averages of word length and to analyse a text by creating what he suggested might be called a 'word-spectrum', or 'characteristic curve', by which he meant a graphic representation of an arrangement of words according to their length and to the relative frequency of their occurrence. Mendenhall (1887: 239) was convinced that

> […] personal peculiarities […] will, *in the long-run*, recur with such regularity that short words, long words, and words of medium length, will occur with definite relative frequencies, so that for him, his approach turned out to be rather an 'application of the doctrine of chance'.

Whereas these early works remained on a merely empirical level, based on intuitive comparisons of visual/graphical impressions, first attempts to develop theoretical models were undertaken in the 1940s and 1950s.

In principle, there are continuous and discrete-frequency models, and it is a matter of philosophy and data structure which kind of approach is favoured, especially since both kinds of model can usually be translated into each other. With regard to word length, discrete models have often been favoured, which is reasonable, since word length is measured in discrete units. The idea of such approaches is to find a mathematical model which, on the basis of observed frequencies, yields theoretical frequencies; these models may have a different number of parameters, and depending on the concrete parameter values—which are estimated by specific methods on the basis of the given empirical data—the final results may vary for one and the same model. The differences between expected and observed values are then submitted to statistical testing. For evaluating the goodness of fit, it is common to apply the X^2 test; since the X^2 value increases with increasing sample size, and the test is therefore increasingly prone to declare differences to be significant, approaches in quantitative linguistics (usually concerned with large sample sizes) prefer to use the standardized determination coefficient $C = X^2/N$.

With regard to word length, the search for theoretical models started in the late 1940s. As far as English is concerned, Elderton's (1949) study deserves mention, in which he suggested the geometric distribution in its 1-shifted form $P_x = p \cdot q^{x-1},\ x = 1,2,3,\ldots$ as an appropriate model. Here, P_x is the probability of a given (word-length) class, p is a parameter to be estimated (with $q = 1–p$). For parameter value $p = 0.7123$ and $q = 1–p = 0.2877$, the theoretical frequencies P_x—represented by white bars in Fig. 2a, alongside the grey bars for the observed frequencies—can be obtained for the above mentioned results of *The Portrait of a Lady*.



Click to view larger

Fig. 2 . Observed (grey bars) and theoretical (white bars) word length distributions (syllables per word) in two texts.

However, the geometric model, with its monotonously decreasing theoretical values, would not be adequate for

other languages, as a comparison with Chekhov's short story 'Dama s sobachkoi' [The Lady with the Dog] shows (see Fig. 2b). Here, the theoretical values are based on the Poisson distribution, first discussed by Russian military doctor S. G. Chebanov (1947), who analysed word-length data from various languages and argued in favour of this model. Like the German physicist Wilhelm Fucks, who later, in a series of works from the 1950s (cf. Fucks 1956), Chebanov was convinced that he had found a universal model. Both Chebanov and Fucks took the 1-parameter Poisson distribution $P_x = \frac{e^{-a} \cdot a^x}{x!}$, $x = 0,1,2,\ldots$ (with parameter *a* to be estimated) as their starting point, displacing it by one since, according to the definition, there were no 0-syllable words, thus obtaining the 1-displaced Poisson distribution $P_x = \frac{e^{-a} \cdot a^{x-1}}{(x-1)!}$, $x = 1,2,3,\ldots$

As can be seen, the fit is very good in both cases, with $C = 0.0054$ and $C = 0.0038$, respectively. Neither model can claim universal relevance for all languages, although early attempts in this field hoped to offer such a perspective; this holds true also for Fucks who used a specific weighted modification of the Poisson distribution, of which the above mentioned 1-displaced model is only a special case (cf. Antić et al. 2005).

In the late 1950s and early 1960s there were some attempts to use the lognormal distribution. Given the characteristic left-skewness of linguistic data (see above), these are assumed to be normally distributed after logarithmic transformation, though such approaches have been mostly abandoned today for theoretical reasons. An important step in the history of word-length modelling, however, was Grotjahn's (1982) suggestion of taking the negative binomial distribution as a standard model which, under specific conditions, converges on the geometric or the Poisson distribution (which thus turn out to be special cases of a more general model). The major impact of this suggestion is not so much the introduction of one more model into the discussion of word length; its importance has instead to be seen in the proposal to concentrate on a variety of distributions which are able to represent a valid 'law of word formation from syllables' (Grotjahn 1982: 73), instead of looking for one general (universal) model.

This idea was subsequently taken up by Grotjahn and Altmann (1993) and elaborated by Wimmer et al. (1994) and Wimmer and Altmann (1996). The basic idea pursued in these papers is that the frequency, or probability, of a given class of *x*-syllable words (P_x) is determined by the class preceding it (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 $P_x = f(x)P_{x-1}$.

Later these ideas, which initially concentrated on word length only, were integrated into Wimmer and Altmann's (2005; 2006) 'Unified derivation of some linguistic laws'. It would lead us too far here to discuss this approach in detail; in short, for a discrete variable *X*, this general approach leads to recurrence formula (1):

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

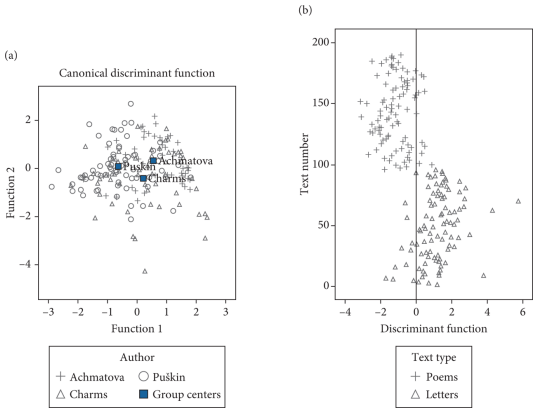
In function (1) we have, in addition to a constant ($1 + a_0$), specific variables (a_i, i=1, 2, …); usually, not more than one or two variables are needed in linguistic modelling. Depending on the exact form of these parameters, different models can be derived: for example, with $a_0 = -1$ and $a_i = 0$ for $i = 2,3,\ldots$ we obtain $P_x = a/x\ P_{x-1}$, which corresponds to the Poisson distribution, and with $-1 < a_0 < 0$ and $a_i = 0$ for $i = 1,2,\ldots$, one obtains the geometric distribution; similarly, most distribution models relevant for linguistics can be derived from this function.

Over the last decades, much empirical evidence has been gathered corroborating hypotheses deduced from this approach. The approach provides a basis for deductive reasoning in quantitative linguistics, and it allows for the derivation of most frequency distributions known in the field of linguistics, word spectra being but one.[4]

With regard to function (1), it can be considered to be a matter of boundary conditions how many and which parameters are needed in a specific research situation. In this respect, individual languages, authorship and personal style, genre, or other factors may be interpreted to represent specific boundary conditions of a general law. In any case, there will not only be cross-linguistic (interlingual) differences; one will always be concerned with intralingual and intertextual (e.g. author or genre specific) differences, too. Even single texts can be shown to be composed of different registers (e.g. narrative or descriptive passages vs. dialogue) ultimately being characterized

by intratextual heterogeneities (Grzybek 2013a).

In practice, it may be a matter of research interest and data situation if models are searched for each individual data set, or if a single model is searched to cover heterogeneous (sub)sets under a common theoretical roof. In this context, it may be appropriate to use mixtures of two distributions, to introduce local modifications (e.g. separate modelling of one-syllable words), or to work with generalizations (against which the individual models converge, or of which they are special cases, under specific circumstances), the more so since there could well be linguistic reasons and justification for such procedures.



Click to view larger

Fig. 3 . Discrimination analyses for 190 letters and poems by three Russian authors.

From what we know today, word length depends on both intratextual and intralingual factors, and it is a matter of academic decision to focus on existing sub-populations as specific sets in their own right, or as variations from a more general, language-specific profile. In any case, individual author-specific factors seem to play only a minor, subordinate role. By way of an example, Fig. 3 shows the result of discrimination analyses, based on 190 Russian texts, a balanced set of letters and poems by three Russian authors (Anna Akhmatova, Daniil Kharms, and Aleksandr Pushkin). Whereas a classification based on authorship yields a poor 38 per cent of correctly discriminated texts (Fig. 3a), a genre-based discrimination improves to 89.5 per cent (Fig. 3b). Obviously, individual variation is relatively limited within genre norms (Kelih et al. 2005).

In summary, it is text type which is a decisive factor influencing word length; as soon as an author enters some textual space, word length is predominantly influenced by basic discourse types, rather than by author-specific factors. The discourse types are not to be identified with functional styles or registers, but are of a more general kind, along distinctions such as dialogical vs. narrative, private vs. official, or oral vs. written. As shown by discrimination analyses of 398 Slovenian and 613 Russian texts from different text types, the best results (92.7 per cent) were obtained for three discourse types: private/oral, public/written, and poetic (Friedl 2006; Grzybek and Kelih 2006).

Across languages, word length of course depends on other linguistic factors too, such as phoneme inventory size, syllable and morphological structure, and degree of analyticity/syntheticity. Some of these factors represent a kind of starting condition, others can be considered to play a crucial role in the given language's processes of self-regulation.

5.2 Word-length Relations

In the previous section word length was treated in a 'self-contained' manner, the textual environment being considered as a kind of global boundary condition. A different perspective is offered when the length of a word is analysed in its direct or indirect relation to other linguistic entities. After all, a word and its length are not isolated phenomena, and any word-length frequency distribution is the product of words' dynamic interactions with other entities in the process of speech generation. In this respect, the following kinds of approach may be distinguished:

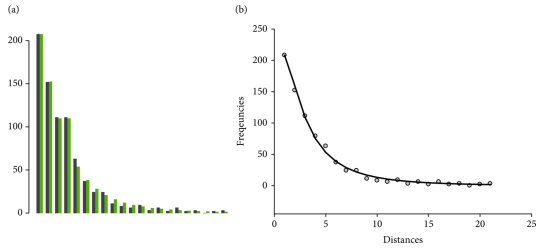
a. *Sequential analyses*. In the simplest case, the length of a word is related to the length of neighbouring words; issues include the distances between words of the same length, word-length *n*-grams, and *L*-motifs

(discussed below).
b. *Positional analyses*. A related though essentially different approach concerns positional aspects of word length in the course of longer text passages, starting from (parts of) sentences up to whole texts.
c. *Horizontal* or *collateral relations*. This perspective takes into consideration the relation of word length to other properties of the word, such as frequency, polysemy, and polytextuality.
d. *Vertical* or *hierarchical* relations. A fourth approach refers to linguistic entities from other structural levels.

5.3 Sequential Analyses

Whereas descriptive characteristics—offering global summarizing measures—and distributional approaches—attempting to describe and model a given sample as a whole—both focus on the linguistic material as a given product, various methods try to take account of procedural aspects, analysing the linguistic data not as some 'given' totality, but—understanding text as a linear sequence of events—in the course of their appearance within the text. These dynamic approaches are here termed 'sequential analyses'.

5.3.1 Word-length Distances



Click to view larger

Fig. 4 . Fitting the Zipf–Alekseev model to word-length distances.

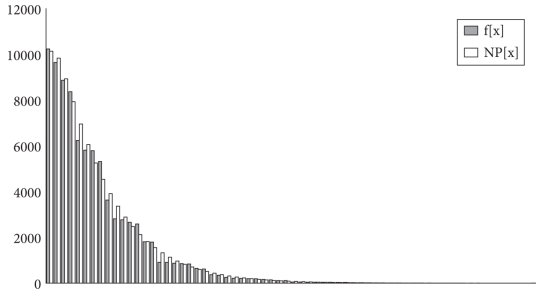
Zörnig (2013a; 2013b) studied the regularity of distances between words of equal length. Defining a real text as a sequence $S = (s_1 ... s_n)$ of length *n*, consisting of elements chosen from the set $\{1, ..., m\}$, where the element *r* occurs exactly k_r times for $r = 1, ..., m (k_1 + ... + k_m = n)$, the distance between two consecutive elements of type *r* is defined as the number of elements $\neq r$ lying between them. Based on the number of occurrences of the distance *d* between two consecutive elements of type *r*, a frequency distribution is obtained, for which Zörnig (2013a) suggests a discrete model or, alternatively, a continuous function (2013b). Among the texts that Zörnig tested was Nikolai Ostrovsky's 1932 novel *Kak zakalialas' stal'* [How the Steel was Tempered]. As Fig. 4 (with distances on the *x*-axis and their frequencies on thy *y*-axis) shows, the Zipf–Alekseev model (cf. Wimmer and Altmann 1999: 665f.) fits both the discrete ($C = 0.17$) and continuous modelling ($R^2 > 0.99$).

5.3.2 Word-length *n*-grams

The concept of *n*-grams, widely used in the fields of computational linguistics and probability, are contiguous sequence of *n* items from a given text; items usually analysed are letters, phonemes, syllables, or words: an *n*-gram of size 1 is referred to as a 'unigram', size 2 is a 'bigram', size 3 is a 'trigram', etc.

Applying this concept to word-length studies, Grzybek and Kelih (2005) analysed the frequency of word-length bigrams: given a sequence 1-3-4-2-5-3-4-1-3-4, for example, we can identify nine pairs 1-3, 3-4, 4-2, 2-5, 5-3, 3-4, 4-1, 1-3, 3-4, of which one (1-3) occurs twice, and another (3-4) three times. For a given text one thus obtains a frequency distribution of length bigrams, which may be rearranged in decreasing order to obtain a rank frequency distribution for which a theoretical distribution model may be searched.

Again, linguistic decisions must be made, such as whether to take sentence boundaries into account; thus far, no systematic studies are available on this matter. As a starting point, in Grzybek and Kelih's (2005) study, ten texts by the Russian author Viktor Pelevin were first submitted to unigram analyses, as described above: for this condition, excellent results were obtained, showing the hyper-Poisson distribution, well-known in quantitative linguistics in general and in word-length studies in particular. Subsequent analysis of the bigram rank frequency distributions showed that they seem to follow a clearly regulated organization, the (right-truncated) negative binomial distribution proving an adequate model in this regard.



Click to view larger

Fig. 5 . Word length bigrams in a Russian text, with observed (grey, f_x) and theoretical (white, NP_x) frequencies.

Fig. 5 represents the results for one of the texts, the novel *Chapaev i pustota* [Chapayev and Void[5]]; the observed frequencies are presented in grey, the theoretical values in white.

5.3.3 Word-length Motifs

Yet another kind of sequential analysis has been suggested by Köhler (2006; 2008) and Köhler and Naumann (2008), who studied groups of word lengths which they term 'motifs'. Köhler defines a length motif (*L*-motif) as the longest continuous series of units (e.g. morphs, words, sentences) of equal or increasing length. In terms of word length measured in syllables, the sentence *Word length studies are almost exclusively devoted to the study of distributions* gives a sequence of five *L*-motifs: (1-1-2)(1-2-4)(3)(1-1-2)(1-4). Second-order *LL*-motifs can be derived. In the above example, there are two *L*-motifs of length 3, followed by one of length 1, etc., resulting in the *LL*-motif sequence (3-3)(1-3)(2). As Köhler shows, the frequency of these motifs can be modelled with distributions well-known in linguistics, such as the Zipf–Mandelbrot or the hyper-Pascal distributions (Wimmer and Altmann 1999: 279ff., 666)

5.4 Positional Analyses

5.4.1 Word-length Dynamics in Running Sentences

Words of a given length are not equally distributed within a sentence; instead, average length tends to increase from beginning to end (Fan et al. 2010; Niemikorpi 1991; 1997; Uhlířová 1997a; 1997b). A reasonable explanation of this phenomenon refers to information theory and theme–rheme (or topic–comment) approaches, implying that in the course of a sentence, new information follows (references to) known information; this explanation would be in line with the well-established fact that longer (and more rarely occurring) words contain more information. A still outstanding question is whether this tendency applies to the intermediate level of clauses and phrases, and, eventually, to the position of a clause within a sentence.

5.4.2 Word-length Dynamics in Running Text

Given the hypothesized increase of information in the course of text segments such as sentences, it is reasonable to assume that the same tendency will characterize larger text segments, or even texts as a whole. (Note that this question makes no sense with regard to text mixtures, or corpora.) In order to test the hypothesis, mean word length must be calculated separately for each sentence (or paragraph, chapter, or text blocks of equal size), and then studied progressively over the course of the text. To date, there is not much empirical evidence concerning these questions. Mention might be made, however, of Kelih's (2012) study of a Russian text (Mikhail Bulgakov's novel *The Master and Margarita*) and its Bulgarian translation, which was indeed able to demonstrate regular tendencies between text and word length. Word length was calculated cumulatively for all 33 chapters, starting with chapter 1, then for chapters 1+2, 1 … 3, 1 … 4, etc., up to the whole novel. Analysing both word-form types and word-form tokens (measured either in the number of graphemes or syllables per word), the hypothesis could be confirmed, the increase of word length (*WoL*) with an increase of text length (*TeL*) being modelled by the simple regression function $WoL = a \cdot TeL^b$. However, the regular increase could only be observed when text length was measured in the number of word-form types, not of word-form tokens.

5.5 Interim Summary

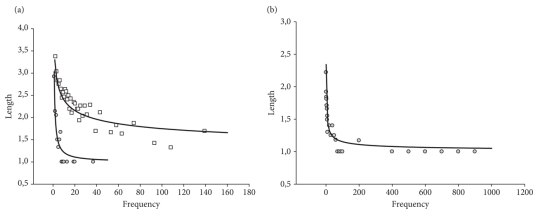
The above list of approaches is not exhaustive; regularities in the organization of word length may be approached in many other ways. In any case, static and dynamic approaches, as outlined above, are not mutually exclusive but complementary: descriptive statistics are simply focused perspectives on a given frequency distribution, and theoretical frequency models not only provide evidence that frequency behaviour as a whole is regularly organized, but also predict the probability of an element of that distribution to occur in the given material, without making prognoses as to when exactly (i.e. at which position) this element is likely to occur. Dynamic approaches, in comparison, provide evidence that sequential order is not randomly organized, but follows particular rules, too. We are still far from understanding the mechanisms in detail; it is however noteworthy that Köhler's results on motifs are strikingly similar to quantitative analyses of syntactic structures, which show a comparable frequency behaviour (Köhler 2012). Similarly, recent research on prose rhythm, concentrating on the distribution of accent and stress in running texts, seems to indicate convergences between word length and rhythmic patterns, insofar as the frequency distribution of distances between stressed/accented syllables appears to be related to the frequency distribution of word-length classes, depending again on the definition of word that is applied, the phonological word being of particular relevance in this context (Grzybek 2013d; 2013e).

5.6 Horizontal/collateral Relations

Collateral or *horizontal* relations concern relations of word length to other properties of the word, such as frequency, polysemy, or polytextuality.

5.6.1 Word Frequency □ Word Length

The relation between word length and word frequency is well known, and has been redundantly corroborated since G. K. Zipf (1935) formulated the corresponding hypothesis (cf. Grzybek and Altmann 2002). Although many details still remain a matter of discussion, it is generally agreed that the more frequently a word is used, the shorter it tends to be; here, too, it is important whether lemmas or word forms are analysed. This is not the place for an extensive discussion of word-frequency issues (for a recent survey of methods, see Popescu et al. 2009); suffice it to say that different word classes (and, as a consequence, their length) may be differently affected by frequency, the distinction between synsemantic and autosemantic (function and content) words being of special importance.

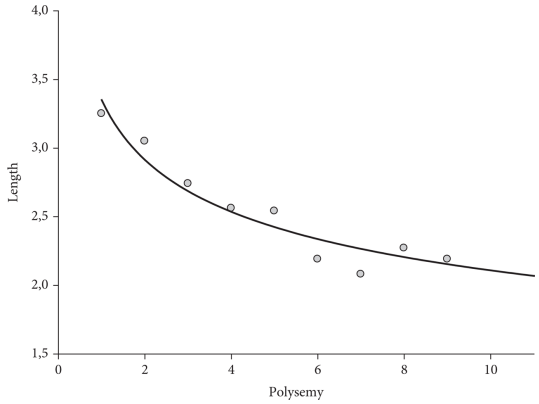


Click to view larger

Fig. 6 . Dependence of word (forms) length on frequency.

Paying special attention to the factors of sample size (or text length) and data homogeneity, Strauss et al. (2006), in their analysis of texts from various languages, found the relation between word length (*WoL*) and frequency (*WoF*) to follow the potency function $WoL = a \cdot WoF^{-b} + 1$. Fig. 6a shows the results for Tolstoy's *Anna Karenina*, both for the first chapter (represented by grey circles) and for all 34 chapters of the first book (white squares); Fig. 6b shows the results for *The Portrait of a Lady*. On the *x*-axis we see the absolute frequency of occurrence in the given text, on the *y*-axis the corresponding word length (in syllables). In both cases, data have been pooled to show the trend more clearly.[6]

5.6.2 Polysemy □ Word Length

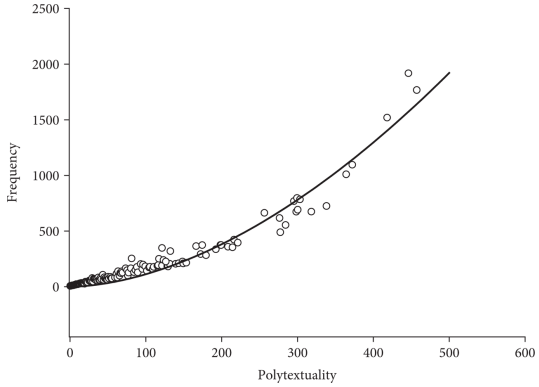


Click to view larger

Fig. 7 . Word length and polysemy.

The relation between polysemy and word length is a repeatedly discussed issue in quantitative linguistics. The direction of dependence has been controversial, both directions in principle being open to testing. Since word prolongation (by affixation, compounding, reduplication, etc.) results from semantic needs (specification or diversification of meaning), one might consider polysemy to be the independent and length the dependent variable; this would result in the assumption that the fewer meanings a word has, the longer it should be, and the more meanings it has the shorter it should be. If, however, shortening is considered to be primarily a result of increased frequency, it seems rather that polysemy should be considered a function of length, shorter words being more likely to be polysemous than longer words. It seems reasonable to side with Köhler (1999), for whom increase of length and decrease of polysemy are simultaneous results of one and the same process. Fig. 7 represents Köhler's results for Māori (based on the analysis of lexematic dictionary material), with length (measured in the number of syllables per lexeme) as the dependent variable.

5.6.3 Polytextuality □ Word Frequency

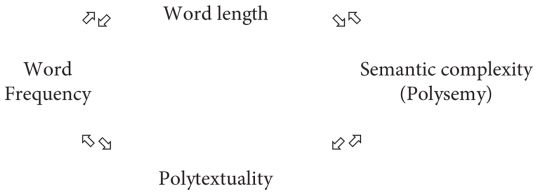


Click to view larger

Fig. 8 . Word frequency and polytextuality.

There is a lawful relationship between the number different environments (or environment types) in which a word occurs (i.e. its polytextuality) and the overall frequency of the word (Köhler 1986; 2006). Since polytextuality—usually measured in terms of the number of different texts in a corpus which contain at least one token of the given word—is related to frequency, and frequency to length, we have an indirect relation between polytextuality and length. Fig. 8 shows the results presented by Köhler (1986), based on an analysis of the German LIMAS corpus:[7] on the *x*-axis we see the number of different texts in which a given word (form) occurs (i.e. its polytextuality), pooled for a given class, on the *y*-axis their frequency of occurrence.

5.6.4 Synergetics: Word Length and Collateral Relations

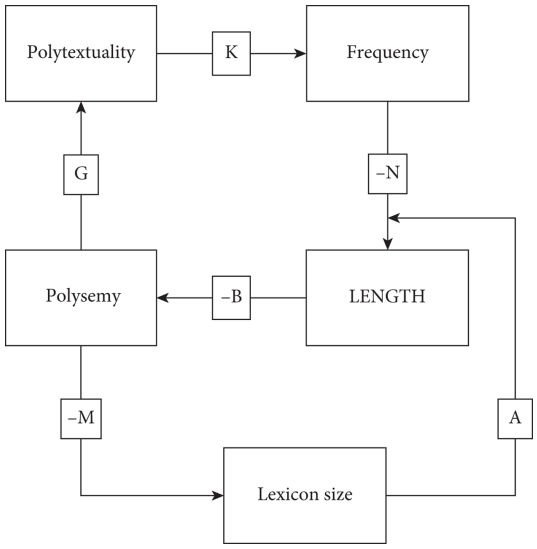


Click to view larger

Fig. 9 . Collateral relations.

On the basis of the factors discussed above which directly or indirectly influence word length, we obtain the following logical chain of reasoning. Frequently used words tend to be shortened; shorter words tend to be polysemous which in turn are likely to be used in more different (con)texts and thus used more frequently. As a result, a system of collateral interrelations emerges, which can be represented in form of a simplified control cycle (Fig. 9).

The cycle in Fig. 9 can be regarded as a small component of a complex synergetic system of linguistic self-regulation (Köhler 2005). Concentrating on the needs of a given system (and, ultimately, its users), synergetics is function- and process-oriented, aiming at functional explanations of dynamic systems through studying processes of self-organization and self-regulation. Synergetic linguistics thus deals with needs and requirements like minimization of production effort (*minP*), memory effort (*minM*), and decoding effort (*minD*), amongst others; since at least some of these are antagonistic by nature—minimal effort for a producer, for example, implies maximal effort for the recipient—the system is in a permanent process of change and dynamic balance, guaranteeing the system's functioning (and, in the successful case, its efficiency and survival).



Click to view larger

Fig. 10 . Partial synergetic model of word length relations.

The relations depicted in Fig. 10 represent a part of the lexical subsystem of a synergetic model of language.

In this schema, rectangles represent system variables (state and control variables), squares represent operators, and arrows stand for effects or bonds; the squares contain symbols for operator types, in our case proportionality operators, with '+' or '–' for their values. For an interpretation of this diagram, one must bear in mind that the original hypotheses have been linearized by way of a logarithmic transformation; that is, in order to interpret the schema, one must use the anti-logarithm along with rules of operator algebra and graph theory. The schema in Fig. 10 thus graphically presents the following hypotheses[8]:

$LS = PS^{-M}$	Lexicon size (LS) is a function of mean polysemy (PS). 'The more polysemous words there are, the smaller the lexicon.'
$WoL = LS^{A}Frq^{-N}$	Word length (WoL) is a function of lexicon size (LS) and frequency (Frq). 'The more words are needed, the longer they are on average; the more frequently a word is used, the shorter it tends to be.'
$PS = WoL^{-B}$	Polysemy (PS) is a function of word length (WoL). 'The longer a word, the less its polysemy.'
$PT = PS^{G}$	Polytextuality (PT) is a function of polysemy (PS). 'Words with high polysemy occur in more different (con)texts.'
$WoF = PT^{K}$	The frequency of a lexical item (WoF) is a function of polytextuality. 'A word is more frequent when it occurs in more different (con)texts.'

5.7 Vertical/hierarchical Relations and the Menzerath–Altmann Law

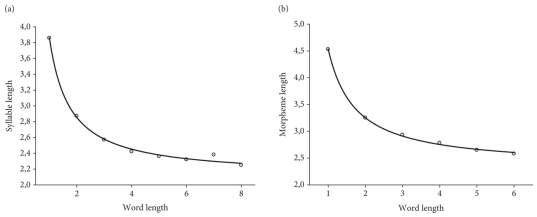
Hierarchical or *vertical* relations concern relations of word length to properties of linguistic entities from other structural levels. Levels, here, are conceived of in classical structuralist terms. On the one hand, we have 'downward' relations of a word to 'lower-level' units such as phone(me)s, letters, or graphemes, these in turn being the constituents of syllables or morphemes; on the other we have 'upward' relations to clauses or phrases, then to sentences, paragraphs, chapters, etc. These levels, and the entities they represent, are of course a matter of linguistic definition. Additional levels may be recognized, depending on text types, e.g. verses or stanzas in poetic texts, sections and books in longer novels. Importantly, all these entities are supposed to have their own regularities, be that with regard to frequency, length, or other properties.

Hierarchical relations can be traced through all structural levels. These relations hold primarily for units of strictly adjacent levels; analyses which leapfrog units from an intervening level are likely to result, not only in an increased degree of variation, but also in more complex, and perhaps even distorted or reverse relations.

In this context, the Menzerath–Altmann Law is of utmost relevance. Generalizing previous findings by Paul Menzerath (1928; 1954) on the relation between word and syllable length, Gabriel Altmann (1980) claimed that, generally, a constituent's length decreases with an increase in the length of the construct; thus, for example, the longer a word, the shorter the syllables which make up the word. This tendency concerns relations between adjacent levels only: the relation between entities from indirectly related levels (e.g. between sentences and words, leapfrogging the intermediate level of clauses or phrases) is expected to show different or even reverse tendencies.

As for intratextual relations, the Menzerath–Altmann Law concerns, first and foremost, the relation of a construct to its immediate constituents. Accordingly, these relations have frequently been modelled with the simple two-parameter function $y = K \cdot x^b$, where *y* represents the construct as the dependent variable, *x* the constituent as the independent variable, *K* some constant, and *b* (for $b < 0$) the steepness of the decrease. This function has long been interpreted as a special case of the more complex function $y = K \cdot x^b \cdot e^{cx}$ (for c = 0), as well as $y = K \cdot e^{cx}$ (for b = 0). More recently, they have all been derived analogically from the more complex function $y = K \cdot e^{ax} \cdot x^b \cdot e^{-c/x}$, which is the continuous equivalent of equation (1), and from which other relevant functions may also be derived. This extension might eventually lead to a partial re-interpretation of previous attempts to find adequate models.

5.7.1 Word Length □ Syllable/morpheme Length



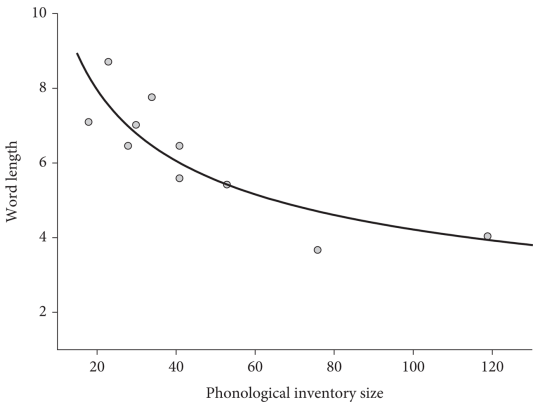
Click to view larger

Fig. 11 . Relation between word length and syllable/morpheme length (based on Menzerath 1954 and Gerlach 1982).

With respect to 'downward' relations of word length, we are concerned with syllable and morpheme structures, these being the direct constituents of the word. In terms of the Menzerath–Altmann Law, syllable/morpheme length is expected to decrease with an increase of word length (measured in the number of syllables, or morphemes, per word). Much empirical corroboration has been gathered in this respect over the last decades. Fig. 11 shows two selected examples from German. Fig. 11a displays data for the word–syllable relation, taken from Menzerath (1954); Fig. 11b illustrates data for the word–morpheme relation from Gerlach (1982); for further illustrations see Altmann and Schwibbe (1989) and Cramer (2005). To model the relation, instead of the potency function $SyL = K \cdot WoL^{-b}$ the exponential function $SyL = K \cdot e^{-c/WoL}$ has been taken here, with $R^2 > 0.99$ in both cases (with $K = 2.11$, $c = 0.60$, and $K = 2.33$, $c = 0.66$, respectively).

5.7.2 Word Length Phoneme Inventory Size

On the next level we are concerned with phonemes, or phonological segments, and related elements. Studies analysing the relation between a language's phoneme inventory size (IS_P) and average word length deserve special mention. Nettle (1995) analysed 50 randomly chosen dictionary entries from ten languages. Referring to Köhler's (1986) theoretical discussions of phonological inventory size as an important factor in the self-regulating processes of language, Nettle concluded that word length is inversely related to the size of phonological inventories, the latter being defined as the number of phonological segments available, with tones being multiplied by the number of vowels.



Click to view larger

Fig. 12 . Relation between phonological inventory size and word length (in phonological segments). Based on Nettle (1995).

Nettle's (1995) results are shown in Fig. 12; included in the figure is the theoretical curve, based on the function $WoL = a \cdot {IS_P}^{-b}$, as used by Nettle. Nettle's (1995) study was extended by Nettle's (1998) analysis of twelve West African languages, and has recently been placed on a wider language basis by Wichmann et al. (2011).

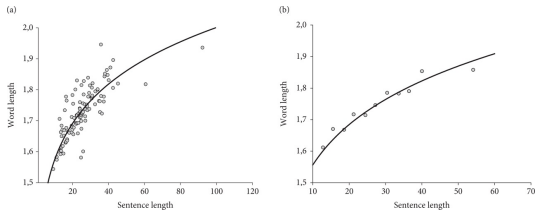
As Fig. 12 shows, the results appear to be convincing; with parameter values $a = 26$ and $b = 0.40$, the fit is $R^2 = 0.72$. The results should, however, be interpreted with caution, because of a number of open questions, some of which have not yet been systematically taken into account in approaches to the length–inventory issue (cf. Kelih 2008; 2010; 2012). Apart from definitional problems (such as the definition of phoneme, or phonological segment, the treatment of tones, or the notion of word analysed in terms of lemmatized dictionary entries), the most problematic issue from a theoretical perspective concerns the presumed direct relation between inventory size

and word length. First, word length has been measured in the number of phonemes (or phonological segments), thus leapfrogging the intermediate level of syllables or morphemes. Second, not only phonological but also phonotactic issues must be taken into consideration, as well as questions of syllable and morpheme structure. For example, it is evident that the more phonemes there are available in a language, the greater the number of different syllable types that can be formed on their basis; this in turn allows for more variation, likely to result in shorter syllables. Shorter syllables, however, are likely to correlate with longer words (measured in the number of syllables per word), following the Menzerath–Altmann Law. The situation is even more complex if phonotactics are taken into account; the more phonemes a language has, the fewer of all possible combinations are actually realised. The situation becomes even more complex when account is taken of frequency of occurrence, not only of phoneme combinations but also of individual phonemes. Of particular relevance here is the proportion of vowels (in their essential syllable-forming function) in the inventory. Despite the disputed assumption that languages with larger phoneme inventories contain relatively fewer vowels, languages with relatively more consonants tend to form more complex syllables, resulting in shorter words (in terms of the number of syllables per word), particularly if frequency of occurrence is taken into consideration.

5.7.3 Word Length □ Clause Length □ Sentence Length

Following the assumption that the Menzerath–Altmann Law also regulates the relationship between the lexical and the sentence level, one might be tempted to expect a decrease of word length with an increase of sentence length. However, this hypothesis would not take into account the intermediate level of clauses, or phrases,[9] which has repeatedly been shown to play an important role in the syntactic processes of self-regulation. In fact, there is abundant evidence proving a regular relation between sentence length and clause length, an increase of sentence length accompanying a decrease of clause length, the latter being measured in the number of words per clause. One should therefore expect an increase of word length with a decrease of clause length and, as a logical consequence, an increase of word length with an increase of sentence length (accompanied by a large portion of variation, due to leapfrogging one analytical level).

Surprisingly, however, the word–clause relation has not to date been empirically studied (Cramer 2005: 672)—a research gap soon to be filled (Grzybek and Rovenchak 2014). What are available, however, are studies on the relation between word length and sentence length, from which eventually indirect evidence can be derived, given the considerations outlined above. However, in this respect, due attention must be paid to the distinction between intratextual and intertextual word–sentence relations.



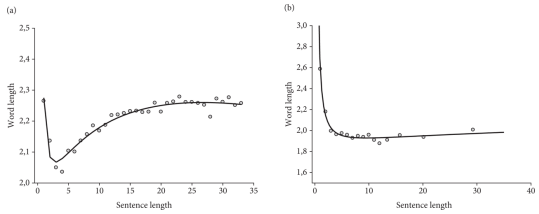
Click to view larger

Fig. 13 . Intertextual relation between sentence length (in words) and word length (in syllables).

The intertextual perspective concerns the study of a sample of texts. For each text, mean word length and average sentence length is calculated separately, the resulting vector of means then being submitted to analysis. Based on results on excerpts from 117 German literary prose texts provided by Arens (1965), Altmann (1983) formulated the Arens–Altmann Law, according to which this vector can be grasped in analogy to the Menzerath–Altmann Law (cf. Grzybek 2013c), which had originally been designed for intratextual relations. Fig. 13a shows the results of fitting to the original data, while Fig. 13b shows the results for pooling sentences in intervals of 3, yielding a significantly better fit.

Recent research has yielded evidence, however, that the state of affairs may be more complex and less clear than hitherto assumed. Separately analysing the relation in homogeneous text types (private letters, dialogues from dramas, short stories, etc.), Grzybek et al. (2007) and Grzybek and Stadlober (2007) found that within each of these genres there is much less variation of word length as compared to sentence length, resulting in a lack of the predicted tendency; in contrast, literary texts, especially novels, seem to be composed of heterogeneous elements, each with its own regulating regime, the overall picture being more the reflex of these different regimes

than of a general rule. As a consequence, the Arens–Altmann Law might turn out to be predominantly valid for the characterization of heterogeneous texts, or text types.



Click to view larger

Fig. 14 . Relation between sentence length (in words) and word length (in syllables) in a Russian and a Slovene text.

Compared to this, the intratextual relation between sentence length and word length has been studied to a lesser degree, for both theoretical and empirical reasons: calculating sentence length in the number of words per sentence leapfrogs the intermediate level of clauses, and measuring in terms of 'indirect units' should not only be avoided in principle, but has also led to results which turned out to be too complex to be grasped by one of the original three versions of the Menzerath–Altmann Law. It was only recently that the sentence–word relation has been modelled for Tolstoy's *Anna Karenina* with the complex function $y = K \cdot e^{ax} \cdot x^b \cdot e^{-c/x}$ mentioned above, resulting in a very good $R^2 = 0.92$. Taking into account that this complex novel consists of heterogeneous text passages (descriptive, narrative, dialogical), and that the immediate sentence–clause relation is leapfrogged, the need for a four-parameter model instead of the less complex potency function seems fully reasonable. Fig. 14a shows the curve for pooled data up to sentence length 30.

By way of a comparison, Figure 14b shows equivalent results for a 1991 Slovene novel, *Zbiralci nasmehov*, by Marijan Pušavec. As can be seen, the trend is also regular, but less complex and in the opposite direction; it seems likely that this reverse tendency is due less to genre factors that to syntactic specifics of Slovene—a hitherto unsolved question requiring systematic study.[10]

5.8 Word Length in an Evolutionary Perspective

Language, as a dynamic system, undergoes evolution—a fact which has long been ignored in linguistics due to the dominating Saussurian dichotomy of synchrony vs. diachrony. Any synchronic cut in a language's history is but an abstract temporal model of the language, and the same is true for any diachronic perspective, which is also but a momentary snapshot at a given historical point of time. And no matter how many snapshots are piled on one another, the result will always be an additive compilation of layers; only modelling the evolving transitions between diachronic cuts, each static by nature, can provide a dynamic understanding of language as an evolving system.

Word length is but one aspect in an overall evolutionary process. The foregoing discussion will have shown that word length, far from being an isolated category in a language or text, is closely interrelated with other linguistic units and levels and forms part of a complex system of interrelations and control cycles. It is evident, therefore, that changes in word length will be related to other changes in the linguistic system. Either change in word length will lead to changes in other elements or it is likely to be the result of other changes.

On an evolutionary perspective on word length, we are concerned with a dynamic (sub)system which, by definition, is subject to change and variability, part of this variability being likely to include processes of diversification. Depending on the perspective taken, diversification may be understood either as a process or as a result of a process. A frequency distribution, for example, may be interpreted as the diachronic result of a previous diversification process (given e.g. an evolutionary process from one-syllable to multi-syllable words), or it may be related to other (simultaneously existing) frequency distributions, which in sum represent an ongoing diversification process (whether stylistic, dialectal, sociolectal, etc.).

Given the complex synergetic embedding of word length in a language's complex dynamic system, it is clear why relevant studies to date have yielded either weak evidence of clear trends, or even contradictory results. Whereas for English, Liberman (2011) observed only a minor decrease in word length in the public speeches of American presidents over a period of about 200 years, Bochkarev et al. (2012) observed an increase over the same period

for British and American texts, as well as for Russian, although with specific fluctuations over the given time period. Whereas these studies were based on letter-counts, Ammermann's (2001) analysis of German letters over a 500-year period was syllable-based. Yet he too found no clear trend, but rather wave-like fluctuations. However, none of these studies controlled relations to other linguistic units which may have locally influenced word length—not only changes in patterns of word formation (derivation, compounding, etc.), but also genre-specific developments, involving changes in sentence length and related factors. As a result, word-length data presently available provide only localized insights into evolutionary questions, and more systematically designed studies in this direction are needed for the theoretical modelling of evolutionary processes of word length.

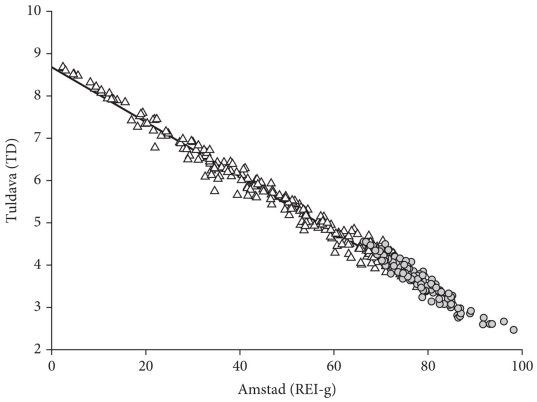
6 Practical Aspects

One of the earliest practical applications of word-length studies was in the sphere of authorship attribution. Indeed, at the very beginning of word-length studies, Mendenhall (see above) addressed the question of unknown or uncertain authorship. It soon turned out, however, that word length, or rather word length alone, is not an appropriate factor for solving authorship issues. Nevertheless, word length continues to be one factor which is taken into account in authorship studies still today, although, more often than not, with insufficient attention to interfering factors such as text typology and other factors like those discussed above.

In fact, word length has traditionally played a particular role for text typological issues, assuming that different text types are characterized by different word length. It is important, of course, what kind of text typology is used (or searched for): since word length covers only a relatively restricted range of variation, within a given language, no typology comprising some 4,000 text sorts (cf. Adamczik 1995) may be expected to find its correlate in word-length differences. Concentrating on rather general types (or registers), one can show that word-length differences concern basic discourse types (along distinctions such as 'public/private', 'oral/written', 'narrative/descriptive/dialogical'), rather than traditional functional styles (cf. Grzybek and Kelih 2006, Grzybek et al. 2005).

Another field where word length has played a crucial role is the measurement of text readability, or reading difficulty. Starting from the 1920s, dozens of readability formulae have been developed, a particularly relevant topic for schoolbook and text book compilations. Based on statistical correlations between readers' (intuitive) estimations and text-linguistic criteria, various measures of text difficulty have been suggested. In the early days of readability research, as many linguistic variables as possible were taken into consideration. It soon became clear, however, that an increased number of variables does not necessarily yield better results, since many of the variables turned out to be interrelated, ultimately measuring one and the same dimension. The subsequent strategy of reducing the number of variables was additionally fostered by the desire to derive formulae which can be handled with maximum ease in everyday practice. In this context, word length and sentence length have always been major criteria; due to its manifold relations to other linguistic categories, word length contains much more information than on length alone.

One of the best known formulae, still widely used, is Flesch's (1948) Reading Ease Index (REI) from the 1940s. This is a linear function, combining a constant with language-specifically weighted sentence length (*SeL*) and word length (*WoL*): $REI_E = 206.835 - 84.6 \cdot WoL - 1.015 \cdot SeL$. This formula applies to English texts only and must be adapted for other languages. For German, Amstad (1978) has suggested the modification $REI_G = 180 - 58.5 \cdot WoL - SeL$ there are similar adaptations for other languages. These adaptations are language-specific, virtually ruling out the possibility of interlingual comparisons (cf. Grzybek 2010).



Click to view larger

Fig. 15 . Comparison of Amstad's (1978) German *REI* (Reading Ease Index) with Tuldava's (1993a; 1993b) TD (Text Difficulty) measure, for technical (white triangle)s and literary (grey circles) texts (cf. Grzybek 2010: 66).

In this respect, Tuldava's (1993a; 1993b) alternative suggestion for measuring text difficulty (*TD*) might turn out to be useful. His formula is based on the simple multiplication of word length (*WoL*) measured by the number of syllables per word and the logarithm of sentence length (*SeL*) measured by the number of words per sentence: $TD = WoL \cdot \ln(SeL)$. Its language-independence renders this formula appropriate for both intra- and interlingual comparisons. A comparison with the German Flesch adaptation has shown a highly significant correlation, proving its obvious efficiency. Fig. 15 shows the results for 240 German texts, separately for literary and technical texts; comparisons with other languages are presently ongoing.

7 Conclusion

It has been a major concern of this chapter to show that word-length behaviour, far from being chaotic or irregular, displays systematic and regular properties. Moreover, word length is not a peripheral or incidental property of the word or indeed of language in general; rather, at least from a quantitative linguistics point of view, word length stands at the intersection of structural levels and functional dimensions. Word length, with its manifold interrelations with other linguistic units, levels, and structures, provides information that goes well beyond word length alone. For this reason, word length needs to be incorporated into a general theory of language.

References

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

Altmann, Gabriel (1978). Towards a theory of language. *Glottometrika* 1: 1–25.

Altmann, Gabriel (1980). Prolegomena to Menzerath's Law. *Glottometrika* 2: 1–10.

Altmann, Gabriel (1983): H. Arens' 'Verborgene Ordnung' und das Menzerathsche Gesetz. In Manfred Faust et al. (eds), *Allgemeine Sprachwissenschaft, Sprachtypologie und Textlinguistik: Festschrift für Peter Hartmann*. Tübingen: Narr, 31–9.

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

Altmann, Gabriel (2013): Aspects of word length. In Reinhard Köhler and Gabriel Altmann (eds), *Issues in Quantitative Linguistics* 3. Lüdenscheid: RAM, 23–38.

Altmann, Gabriel, and Schwibbe, Michael (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Hildesheim: Olms.

Ammermann, Stefan (2001). *Zur Wortlängenverteilung in deutschen Briefen über einen Zeitraum von 500 Jahren*.

In Karl-Heinz Best (ed.), *Häufigkeitsverteilungen in Texten*. Göttingen: Peust & Gutschmidt, 59–91.

Amstad, Toni (1978). Wie verständlich sind unsere Zeitungen. Dissertation, University of Zürich.

Antić, Gordana, Grzybek, Peter, and Stadlober, Ernst (2005). Mathematical aspects and modifications of Fucks' generalized Poisson distribution. In Reinhard Köhler, Gabriel Altmann, and Raimund G. Piotrovski (eds), *Quantitative Linguistik—Quantitative Linguistics: Ein Internationales Handbuch—An International Handbook*. Berlin: de Gruyter, 158–80.

Arens, Hans (1965). *Verborgene Ordnung: die Beziehungen zwischen Satzlänge und Wortlänge in deutscher Erzählprosa vom Barock bis heute*. Düsseldorf: Schwann.

Best, Karl-Heinz (2005). Wortlängen. In Reinhard Köhler, Gabriel Altmann, and Raimund G. Piotrovski (eds), *Quantitative Linguistik—Quantitative Linguistics: Ein Internationales Handbuch—An International Handbook*. Berlin: de Gruyter, 260–73.

Bochkarev, Vladimir V., Shevlykova, Anna V., and Solovyev, Valery D. (2012). Average word length dynamics as indicator of cultural changes in society. [**http://arxiv.org/abs/1208.6109**]

Chebanov, Sergei G. (1947). On conformity of language structures within the Indo-European family to Poisson's Law. In *Comptes Rendus (Doklady) de l'Académie des Sciences de l'URSS* 55(2): 99–102.

Cramer, Irene (2005). Das Menzerathsche Gesetz. In Reinhard Köhler, Gabriel Altmann, and Raimund G. Piotrowski (eds), *Quantitative Linguistik—Quantitative Linguistics. Ein Internationales Handbuch—An International Handbook*. Berlin: de Gruyter, 650–88.

Dixon, Robert M. W., and Aikhenvald, Alexandra Y. (2002). Word: a typological framework. In: Dixon and Aikhenvald (eds), *Word: A Cross-Linguistic Typology*. Cambridge: Cambridge University Press, 1–41.

Elderton, William P. (1949). A few statistics on the length of English words. *Journal of the Royal Statistical Society*, series A: *General* 112: 436–45.

Fan, Fenxiang, Grzybek, Peter, and Altmann, Gabriel (2010). Dynamics of word length in sentence. *Glottometrics* 20: 70–109.

Flesch, Rudolf (1948). A new readability yardstick. *Journal of Applied Psychology* 32(3): 221–33.

Friedl, Alexander (2006). Untersuchungen zur Texttypologie im Russischen anhand von 609 Texten in 13 Textsorten. MA thesis, Graz University.

Fucks, Wilhelm (1956). Mathematical theory of word formation. In Colin Cherry (ed.), *Information Theory*. New York: Academic Press, 154–70.

Gerlach, Rainer (1982). *Zur Überprüfung des Menzerath'schen Gesetzes im Bereich der Morphologie*. In: *Glottometrika 4*. Bochum: Brockmeyer, 95–102.

Grotjahn, Rüdiger (1982). Ein statistisches Modell für die Verteilung der Wortlänge. *Zeitschrift für Sprachwissenschaft* 1: 44–75.

Grotjahn, Rüdiger, and Altmann, Gabriel (1993). Modeling the distribution of word length: some methodological problems. In Reinhard Köhler and Burkhard B. Rieger (eds), *Contributions to Quantitative Linguistics*. Dordrecht: Kluwer, 141–53.

Grzybek, Peter (2006). History and methodology of word length studies: the state of the art. In Peter Grzybek (ed.), *Contributions to the Science of Text and Language: Word Length Studies and Related Issues*. Dordrecht: Springer, 15–90.

Grzybek, Peter (2007). On the systematic and system-based study of grapheme frequencies: a re-analysis of German letter frequencies. *Glottometrics* 15: 82–91.

Grzybek, Peter (2010). Text difficulty and the Arens–Altmann Law. In Peter Grzybek, Emmerich Kelih, and Ján Mačutek (eds), *Text and Language: Structures, Functions, Interrelations, Quantitative Perspectives*. Vienna: Praesens, 57–70.

Grzybek, Peter (2013a). Homogeneity and heterogeneity within language(s) and text(s): theory and practice of word length modeling. In Reinhard Köhler and Gabriel Altmann (eds), *Issues in Quantitative Linguistics* 3. Lüdenscheid: RAM, 66–99.

Grzybek, Peter (2013b). The emergence of stylometry: prolegomena to the history of term and concept. In Katalin Kroó and Peeter Torop (eds), *Studies in 19th Century Literature*. Budapest: L'Harmattan. 58-75.

Grzybek, Peter (2013c). Arens–Altmann Law. In Reinhard Köhler, Peter Grzybek, and Sven Naumann (eds), *Formale und Quantitative Linguistik*. Berlin: de Gruyter.

Grzybek, Peter (2013d). Empirische Textwissenschaft: Prosarhythmus im ersten Drittel des 20. Jahrhunderts als historisch-systematische Fallstudie. In Aage Hansen-Löve, Brigitte Obermayr, and Georg Witte (eds), *Form und Wirkung: Phänomenologische und empirische Kunstwissenschaft in der Sowjetunion der 1920er Jahre*. Munich: Fink, 427–55.

Grzybek, Peter (2013e). Samoreguliatsiia v tekste (na primere ritmicheskikh protsessov v proze). In Igor' A. Pil'shchikov (ed.), *Sluchainost' i nepredskazuemost' v istorii kul'tury*. Tallinn: Acta Universitatis Tallinniensis, 78–115.

Grzybek, Peter, and Altmann, Gabriel (2002). Oscillation in the frequency–length relationship. *Glottometrics* 6: 97–107.

Grzybek, Peter, and Kelih, Emmerich (2005). Häufigkeiten von Wortlängen und Wortlängenpaaren: Untersuchungen am Beispiel russischer Texte von Viktor Pelevin. In Eva Binder, Wolfgang Stadler, and Helmut Weinberger (eds), *Zeit—Ort—Erinnerung: Slawistische Erkundungen aus sprach-, literatur- und kulturwissenschaftlicher Perspektive*. Innsbruck: Institut für Sprachen und Literaturen der Universität Innsbruck, 395–407.

Grzybek, Peter, and Kelih, Emmerich (2006). Empirische Textsemiotik und quantitative Text-Typologie. In Jeff Bernard, Jurij Fikfak, and Peter Grzybek (eds), *Text & Reality/ Text & Wirklichkeit*. Ljubljana: ZRC, 95–120.

Grzybek, Peter, Kelih, Emmerich, and Stadlober, Ernst (2009). Slavic letter frequencies: a common discrete model and regular parameter behavior? In Reinhard Köhler (ed.), *Issues in Quantitative Linguistics*. Lüdenscheid: RAM, 17–33.

Grzybek, Peter, and Stadlober, Ernst (2007). Do we have problems with Arens' Law? A new look at the sentence–word relation. In Peter Grzybek and Reinhard Köhler (eds), *Exact Methods in the Study of Language and Text: Dedicated to Gabriel Altmann on the Occasion of his 75th Birthday*. Berlin: Mouton de Gruyter, 205–17.

Grzybek, Peter, Stadlober, Ernst, and Kelih, Emmerich (2007). The relationship of word length and sentence length: the inter-textual perspective. In Reinhold Decker and Hans-J. Lenz (eds), *Advances in Data Analysis*. Berlin: Springer, 611–18.

Grzybek, Peter, Stadlober, Ernst, Kelih, Emmerich, and Antić, Gordana (2005). Quantitative text typology: the impact of word length. In Claus Weihs and Wolfgang Gaul (eds), *Classification: The Ubiquitous Challenge*. Heidelberg: Springer, 53–64.

Kelih, Emmerich (2008). Phoneminventar—Wortlänge: Einige grundsätzliche Überlegungen. *Aktual'ni problemy hermans'koï filolohiï*. Chernivtsi: *Knihi XXI*, 25–9.

Kelih, Emmerich (2010). Wortlänge und Vokal- Konsonantenhäufigkeit: Evidenz aus slowenischen, makedonischen, tschechischen und russischen Paralleltexten. *Anzeiger für Slavische Philologie* 36: 7–27.

Kelih, Emmerich (2012). Systematic interrelations between grapheme frequencies and the word length: empirical evidence from Slovene. *Journal of Quantitative Linguistics* 19(3): 205–31.

Kelih, Emmerich, Antić, Gordana, Grzybek, Peter, and Stadlober, Ernst (2005). Classification of author and/or genre? The impact of word length. In Claus Weihs and Wolfgang Gaul (eds), *Classification: The Ubiquitous Challenge*. Heidelberg: Springer, 498–505.

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

Köhler, Reinhard (1999). *Der Zusammenhang zwischen Lexemlänge und Polysemie im Maori*. In Jozef Genzor and Slavomír Ondrejovič (eds), *Pange lingua*. Bratislava: Veda, 27–34.

Köhler, Reinhard (2005). Synergetic linguistics. In Reinhard Köhler, Gabriel Altmann, and Raimund G. Piotrovski (eds), *Quantitative Linguistik—Quantitative Linguistics. Ein Internationales Handbuch—An International Handbook*. Berlin: de Gruyter, 760–74.

Köhler, Reinhard (2006). The frequency distribution of the lengths of length sequences. In: J. Genzor, and M. Bucková (eds.), *Favete Linguis: Studies in Honour of Viktor Krupa*. Bratislava: Academic Press; 142-52.

Köhler, Reinhard (2008). Sequences of linguistic quantities: Report on a new unit of investigation. *Glottotheory*, 1(1), 115-9.

Köhler, Reinhard (2012). *Quantitative Syntax Analysis*. Berlin: de Gruyter.

Köhler, Reinhard, and Naumann, Sven (2008). Quantitative text analysis using L-, F- and T-segments. In Burkhard Preisach et al. (eds), *Data Analysis: Machine Learning and Applications*. Berlin: Springer; 637–46.

Liberman, Mark (2011). Real trends in word and sentence length. [**http://languagelog.ldc.upenn.edu/nll/?p=3534**]

Mendenhall, Thomas C. (1887). The characteristic curves of composition. *Science* 9(214): 237–46.

Menzerath, Paul (1954). *Die Architektonik des deutschen Wortschatzes*. Bonn: Dümmler.

Menzerath, Paul, and de Oleza, José M. (1928). *Spanische Lautdauer: Eine experimentelle Untersuchung*. Berlin: de Gruyter.

Nettle, Daniel (1995). Segmental inventory size, word length, and communicative efficiency. *Linguistics* 33: 359–68.

Niemikorpi, Antero (1991). Suomen kielen sanaston dynamiikkan. *Acta Wasaensia* 26(2). Vaasa: Vaasan yliopisto.

Niemikorpi, Antero (1997). Equilibrium of words in the Finnish Frequency Dictionary. *Journal of Quantitative Linguistics* 4(1–3): 190–96.

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

Popescu, Ioan-Iovitz, Altmann, Gabriel, Grzybek, Peter, Jayaram, Bijapur D., Köhler, Reinhard, Krupa, Viktor, Mačutek, Ján, Pustet, Regina, Uhlířová, Ludmila, and Vidya, Matummal N. (2009). *Word Frequency Studies*. Berlin: Mouton de Gruyter.

Popescu, Ioan-Iovitzu, Naumann, Sven, Kelih, Emmerich, Rovenchak, Andrij, Sanada, Haruko, Overbeck, Anja, et al. (2013). Word length: aspects and languages. In Reinhard Köhler and Gabriel Altmann (eds), *Issues in Quantitative Linguistics 3. Dedicated to Karl-Heinz Best on the Occasion of his 70th Birthday*. Lüdenscheid: RAM, 224–81.

Strauss, Udo, Grzybek, Peter, and Altmann, Gabriel (2006). Word length and word frequency. In Peter Grzybek (ed.), *Contributions to the Science of Text and Language: Word Length Studies and Related Issues*. Dordrecht: Springer, 277–95.

Tuldava, Juvan (1993a). Measuring text difficulty. In Gabriel Altmann (ed.), *Glottometrika* 14: 69–81.

Tuldava, Juvan (1993b). The statistical structure of a text and its readability. In Luděk Hřebíček and Gabriel Altmann

(eds), *Quantitative Text Analysis*. Trier: WVT, 215–27.

Uhlířová, Ludmila (1997a). O vztahu mezi délkou slova a jeho polohou ve větě. *Slovo a slovesnost* 58: 174–84.

Uhlířová, Ludmila (1997b). Length vs. order: word length and clause length from the perspective of word order. *Journal of Quantitative Linguistics* 4(1–3): 266–75.

Wichmann, Søren, Rama, Taraka, and Holman, Erich W. (2011). Phonological diversity, word length, and population sizes across languages: the ASJP evidence. *Linguistic Typology* 15(2): 177–97.

Wimmer, Gejza, and Altmann, Gabriel (1996). The theory of word length: some results and generalizations. *Glottometrika 15: Issues in General Linguistic Theory and the Theory of Word Length*. Trier: WVT, 112–33.

Wimmer, Gejza, and Altmann, Gabriel (1999). *Thesaurus of Univariate Discrete Probability Distributions*. Essen: Stamm.

Wimmer, Gejza, and Altmann, Gabriel (2005). Unified derivation of some linguistic laws. In Reinhard Köhler, Gabriel Altmann, and Raimund G. Piotrovski (eds), *Quantitative Linguistik—Quantitative Linguistics: Ein Internationales Handbuch—An International Handbook*. Berlin. New York: de Gruyter, 791–807.

Wimmer, Gejza, and Altmann, Gabriel (2006). Towards a unified derivation of some linguistic laws. In Peter Grzybek (ed.), *Contributions to the Science of Text and Language: Word Length Studies and Related Issues*. Dordrecht: Springer, 329–37.

Wimmer, Gejza, Köhler, Reinhard, Grotjahn, Rüdiger, and Altmann, Gabriel (1994). Towards a theory of word length distribution. *Journal of Quantitative Linguistics* 1(1): 98–106.

Zörnig, Peter (2013a). Distances between words of equal length in a text. In Reinhard Köhler and Gabriel Altmann (eds), *Issues in Quantitative Linguistics* 3. Lüdenscheid: RAM, 117–29.

Zörnig, Peter (2013b). A continuous model for the distances between coextensive words in a text. *Glottometrics* 25: 54–68.

Notes:

([1]) With regard to the initial definition of length given above, one may argue against Altmann's (2013) recent suggestion to consider measuring word length on the basis of morphemes as a measure of complexity rather than of length; such different views may be related, however, to the treatment of morphemes, e.g., the *–s* morpheme in the English verb form 'runs', either as one morpheme, or as a complex of four morpheme functions (3rd person, singular, indicative, present tense).

([2]) The concept of mora originates in classical verse theory, where it is understood as the smallest time unit with regard to verse and syllable duration. In modern linguistics, it is defined as a psycho-physiologically perceptible measure, primarily in the fields of phonetics/phonology and prosody research as a measure of syllable weight. The definitions in this field are not cross-linguistically unified. Generally speaking, the following rules hold: a syllable onset (i.e., the first consonant/s of a syllable) is not considered to represent a mora; a syllable nucleus with a short vowel, or a short vowel with maximally one following consonant, constitutes one mora (such syllables being termed 'monomoraic' or 'light'); and syllables with a long vowel or with one short vowel and more than one consonant are counted as two morae and termed 'bimoraic' or 'heavy'. In some languages (e.g., Japanese), the coda (i.e., the consonant/s of a syllable which follow the nucleus) represents one mora, in others not, and for some the state of affairs is unclear. In English, for example, the final consonant of a stressed syllable may be considered to constitute a separate mora; thus the word *cat*, if stressed, would be bimoraic, whereas the identical unstressed syllable in *tomcat* would be monomoraic.

([3]) For the history of word length studies in general, see Grzybek (2006); for the importance of word length in 19th stylistics, see Grzybek (2013b).

([4]) From the equivalent continuous approach, many continuous functions, the relevance of which for linguistics

has repeatedly been proven over the years, can likewise be derived.

([5]) This novel, first published in 1996, is also known in the US as *Buddha's Little Finger*, and in the UK as *Clay Machine Gun*.

([6]) There are different methods of data pooling, which are usually used in case of sparse data. Pooling procedures require careful processing: on the one hand, they serve to make initially hidden structures more clearly visible, on the other hand, pooling must retain exactly these structures and not destroy them.

([7]) The LIMAS corpus (cf. http://korpora.zim.uni-duisburg-essen.de/Limas/index.htm) consists of 500 texts and text passages, each of ca. 2000 word lengths, thus summing up to 1 million words.

([8]) As a matter of fact, the schema and the equations derived from it are gross simplifications, concentrating on those system components discussed above, and omitting additional system requirements and interactions between them, as well as further components to be integrated.

([9]) Again, what counts as a clause is of course a matter of definition, which may change for different languages. Other units, such as phrases, may also be appropriate for this intermediate level between word and sentence.

([10]) In further pursuing such intra-textual (self)regulatory mechanisms, one should be aware of the fact that sentence length, too, is not a "given" unit; rather, there seem to be rule-like relations (again following the Menzerath-Altmann law) between sentence length and supra-sentential units like paragraphs, or chapters, depending on the text type studied (cf. Grzybek 2011, 2013). Taking into account that relating word length to such supra-sentential units is leapfrogging more than one level, no straightforward results should be expected, however.

Peter Grzybek

Peter Grzybek, Institute for Slavic Studies, University of Graz