Quantitative Text Typology
The Impact of Sentence Length

Emmerich Kelih[1], Peter Grzybek[1], Gordana Antić[2], and Ernst Stadlober[2]

[1] Department for Slavic Studies,
University of Graz, A-8010 Graz, Merangasse 70, Austria
[2] Department for Statistics,
Technical University Graz, A-8010 Graz, Steyrergasse 17/IV, Austria

Abstract. This study focuses on the contribution of sentence length for a quantitative text typology. Therefore, 333 Slovenian texts are analyzed with regard to their sentence length. By way of multivariate discriminant analyses it is shown that indeed, a text typology is possible, based on sentence length, only; this typology, however, does not coincide with traditional text classifications, such as, e.g., text sorts or functional style. Rather, a new categorization into specific discourse types seems reaonable.

1 Sentence length and text classification: methodological remarks

Text research, based on quantitative methods, is characterized by two major spheres of interest: (1) quantitative text classification, in general (cf., e.g., Alekseev 1988), and (2) authorship discrimination and attribution of disputed authorship, in particular (cf., e.g., Smith 1983). Both lines of research are closely interrelated and share the common interest to identify and quantify specific text characteristics, with sentence length playing a crucial role and obviously being an important factor. However, in most approaches sentence length is combined with other quantitative measures as, e.g., the proportion of particular parts of speech, word length (usually measured by the number of letters per word), the proportion of specific prepositions, etc. (cf. Karlgren/Cutting 1994, Copeck et al. 2000). This, in fact, causes a major problem, since the specific amount of information, which sentence length may provide for questions of text classification, remains unclear.

The present study starts at this particular point; the objective is an empirical analysis based on a corpus of 333 Slovenian texts. From a methodological perspective, the procedure includes the following steps, before multivariate discriminant analyses will be applied to quantitative text classification:

a. the theoretical discussion of qualitative approaches to text classification, mainly of research in the realm of text sorts and functional styles, and the relevance of these classifications for empirical studies;
b. the elaboration of an operational definition of 'sentence' as well as of a consistent measuring unit;

c. the derivation of adequate statistical characteristics from the frequency distribution of sentence lengths, in addition to average sentence length.

1.1 Definition of 'word' and 'sentence'

In this study, 'sentences' are considered to be constitutive units of texts, separated from each other by punctuation marks; by way of a modification of usual standards, the definition of sentence used in this study is as follows:

Definition 1. The punctuation marks [.], [. . .], [?], and [!] function as sentence borders, unless these characters are followed by a capital letter in the initial position of the subsequent word.

This definition is not claimed to be of general linguistic validity; rather, it turns out to be adequate for our corpus of pre-processed texts, taken from the Graz Quantitative Text Analysis Server (QuanTAS).[1] Now, as far as the measuring unit of sentence length is concerned, often the number of clauses is claimed to be adequate, since clauses are direct constituents of sentences. Yet, in our study, the number of words per sentence is preferred, a word being defined as an orthographic-phonetic unit. Apart from the fact, that we thus have very operational definitions of units at our disposal, control studies including alternative definitions of both 'word' and 'sentence' have shown that both definitions are rather stable, and that a change of definition results in shifts of systematic nature (Antić et al. 2005, Kelih/Grzybek 2005).

1.2 Text basis, methods of classification, and statistical characteristics applied

The 333 Slovenian texts under study, have not been arbitrarily chosen; rather, they were supposed to cover the broad spectrum of possible genres, and thus to be representative for the textual world in its totality. Therefore, the texts were taken from the above-mentioned corpus, in which each text has been submitted to a qualitative a priori classification, according to which each text is attributed to a particular text sort. The theoretical distinction of text sorts being based on specific communicative-situational factors (cf. Adamzik 2000). For the present study, all text sorts have additionally been attributed to functional styles: as opposed to text sorts, the theory of functional styles (cf. Ohnheiser 1999) refers to rather general communicative characteristics. The degree of abstractness is extremely different in case of texts sorts and functional styles: whereas contemporary research in text sorts distinguishes about 4,000 different text sorts, functional styles usually confine to a number of about six to eight. Any kind of qualitative generalization necessarily results

[1] This data base contains ca. 5,000 texts from Croatian, Slovenian, and Russian; all texts are pre-processed and specifically tagged; this procedure guarantees a unified approach.– Cf.: `http://www-gewi.uni-graz.at/quanta`

in some kind of uncertainty relation and may lead to subjective decisions. On the one hand, such subjective decisions may be submitted to empirical testing, attempting to provide some intersubjectively approved agreements (cf. Grzybek/Kelih 2005). On the other hand, one may investigate in how far qualitatively obtained classifications, taken as mere tentative a priori classifications, bear a closer empirical examination.

This paper follows the second direction: our aim is to study, (a) to what degree a classification of texts can be achieved on the basis of sentence length (or, to put it in other words, to what degree sentence length may contribute to a classification of texts), and (b) in how far qualitative classifications involving either (b_1) text sorts or (b_2) functional styles correspond to the empirical findings. Table 1 represents the involved spectrum of text sorts and functional styles, along with a number of statistical characteristics described below. As was mentioned above, in this study each individual text is treated as

Table 1. Text sorts and functional styles: some statistical characteristics

Functional style	Text sort	m_1	s	h	S	total
Everyday style	Private letters	15.40	10.08	3.79	7.55	31
Administrative style	Recipes	10.09	4.39	3.05	3.40	31
	Open Letters	26.07	14.25	4.63	15.66	29
Science	Humanities	21.53	11.71	4.55	22.31	46
	Natural sciences	20.88	11.10	3.75	13.55	32
Journalistic style	articles	23.46	11.18	3.76	8.27	43
	Readers' letters	23.75	13.01	3.98	21.16	30
Literary prose	Novels	14.24	8.48	4.51	4.32	49
Drama	Dramatic texts	6.48	5.38	3.60	13.85	42

a separate object: for each individual text, sentence lengths are measured by the number of words per sentence. Thus, a frequency distribution of x-word sentences is obtained. From this frequency distribution, a set of statistical variables can be derived, such as: mean ($\bar{x} = m_1$), variance ($s^2 = m_2$), standard deviation (s), entropy ($h = -\sum p \cdot \mathrm{ld} p$), the first four central moments (m_1, m_2, m_3, m_4) and quotients, such as the coefficient of variation $v = \bar{x}/s$), Ord's $I = m_2/m_1$ Ord's $S = m_3/m_2$, and many others. This pool of variables – ca. 35 variables have been derived for our analyses (cf. Grzybek et al. 2005) – serves as a basis for multivariate discriminant analyses. Of course, the aim is to use only a minimum of these variables;[2] therefore, the 35 variables are tested for their relevance in text classification in a preliminary study. As a first result, it turns out that there are four dominant characteristics,

[2] The corresponding procedures have proven to be efficient with regard to word length studies by the authors of this text before, and they shall be applied to sentence length studies, here.

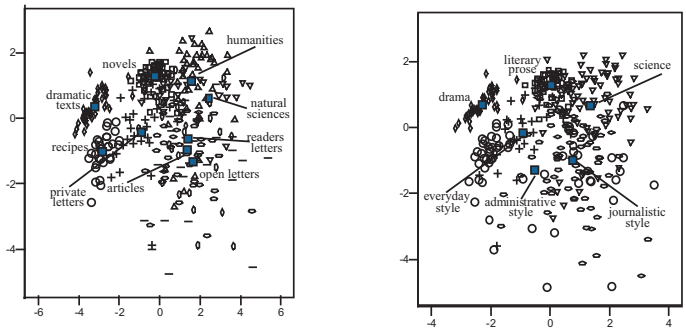
which are important for all subsequent steps: (i) average sentence length $\bar{x}$, (ii) standard deviation s, (iii) Ord's criterion S, and (iv) entropy h.

Notwithstanding the treatment of individual texts, Table 1 offers a general orientation, representing the values of these statistical characteristics with regard to the nine text sorts. Obviously, there are tremendous differences between the various functional styles and, within each functional style, between different text sorts. On the one hand, these observations imply a clear warning as to any corpus-based approach, not paying due attention to genre diversity. On the other hand, these observations give reason to doubt the adequacy of merely qualitative classifications.

2 Sentence length and discriminant analyses

2.1 Submitting the qualitative classifications to empirical testing

On the basis of the above-mentioned discussion, the question arises in how far the tentative a priori attribution of individual texts (a) to text sorts and (b) to functional style is corroborated by sentence length analyses. The results of multivariate discriminant analyses show that only 62.50% of the texts are correctly attributed to one of the nine text sorts; likewise, only 66.40% of the texts are correctly attributed to one of the six functional styles – cf. Fig. 1. This result indicates that neither text sorts nor functional styles can be adequate categories for text classifications based on sentence length.



(a) Text Sorts (b) Functional Styles

Fig. 1. Results of Discriminant Analyses

Taking account of these results, it seems plausible to search for a new type of text classification. This classification should start with text sorts,

since they are more specific than the more general functional styles. Given a number of nine text sorts, a first step in this direction should include the stepwise elimination of individual text sorts.

2.2 Stepwise reduction: temporary elimination of text sorts

An inspection of Fig. 1(a) shows that dramatic texts and cooking recipes cover relatively homogeneous areas in our sample of 333 texts. This is a strong argument in favor of assuming sentence length to be a good discriminating factor for these two texts sorts. Consequently, temporarily eliminating these two text sorts from our analyses, we can gain detailed insight into the impact of sentence length on the remaining seven text sorts (private letters, scientific texts from human and natural sciences, open letters, journalistic articles, readers' letters, and novels). As multivariate discriminant analyses of these remaining 266 texts show, an even less portion of only 51.9% are correctly classified. However, 98% (48 of 49) of our novel texts are correctly classified, followed by the private letters; as to the latter, 64.5% are correctly classified, but 25.8% are misclassified as novels. Obviously, novel texts and private letters seem to have a similar form as to their sentence length; therefore, these two text sorts shall be temporarily eliminated in the next step.

2.3 Stepwise reduction: formation of new text groups

The remaining five text sorts (human and natural sciences, open and readers' letters, articles) consist of 180 texts. Discrimination analyses with these five text sorts lead to the poor result of 40% correct classifications. Yet, the result obtained yields an interesting side-effect, since all text sorts are combined to two major groups: (i) scientific texts, and (ii) open letters and letters. Attributing readers' letters – which almost evenly split into one of these two groups – to the major group of journalistic texts, we thus obtain two major text groups: 78 scientific texts, and 102 journalistic texts. A discriminant analysis with these two groups results in a relatively satisfying percentage of 82.20% correct classifications (cf. Table 2).

Table 2. Attribution of Scientific and Journalistic Texts

Text groups	Group membership		
	Scientific texts	Journalistic texts	total
Scientific texts	65	13	78
Journalistic texts	19	83	102

Since the consecutive elimination of text sorts (recipes, dramatic texts, novel texts, private letters) has revealed that the remaining five text sorts

form two global text groups, the next step should include the stepwise re-introduction of all temporarily eliminated text sorts.

3 Re-integration: towards a new text typology

Re-introducing the previously eliminated text sorts, particular attention has to be paid to the degree of correct classification, the percentage of 82.2% obtained above representing some kind of benchmark. In detail, the following percentages were obtained:

1. re-introducing the cooking recipes (three major texts groups, $n = 211$) results in 82.5% correct classifications;
2. additionally re-integrating the dramatic texts (four major texts groups, $n = 253$) even increases the percentage of correct classifications to 86.60%;
3. also re-integrating the novel texts (five major texts groups, $n = 302$) still results in 82.5% correct classifications;
4. finally re-integrating the last missing text sort (private letters) finally yields 78.8% correct classifications of $n = 333$ texts (cf. Table 3).

Table 3. Six Text Groups

Text group	CR	ST	PL	JT	NT	DT	total
Cooking Recipes (CR)	30	0	0	0	0	1	31
Scientific Texts (ST)	0	58	0	11	9	0	78
Private Letters (PL)	3	1	16	2	8	1	31
Journalistic Texts (JT)	0	16	7	71	8	0	102
Novel Texts (NT)	0	2	0	1	46	0	49
Dramatic Texts (DT)	0	0	0	0	2	40	42

The synoptical survey of our new classification allows for a number of qualitative interpretations: Obviously, sentence length is a good discriminant for dramatic texts, probably representing oral speech in general (in its fictional form, here). The same holds true for the very homogeneous group of cooking recipes, most likely representing technical language, in general. Sentence length also turns out to be a good discriminating factor for novel texts with a percentage of ca. 94% correct classifications. Scientific texts and journalistic texts form two major groups which are clearly worse classified as compared to the results above (74.35% and 69.61%, respectively); however, the majority of mis-classifications concern attributions to the opposite group, rather than transitions to any other group.

As compared to this, private letters –which were re-introduced in the last step – represent a relatively heterogeneous group: only 51.61% are correctly classified, 25.81% being attributed to the group of novel texts.

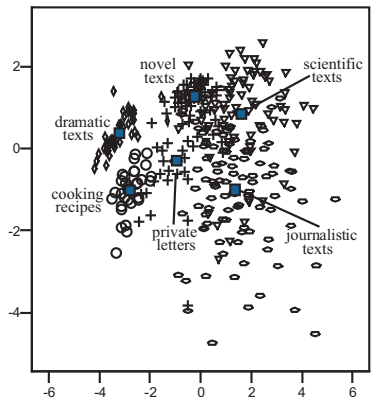


Fig. 2. Discriminant Analysis of Six Text Groups

4 Summary

The present study is a first systematic approach to the problem of text classification on the basis of sentence length as a decisive discriminating factor. The following major results were obtained:

a. Taking the concept of functional styles as a classificatory basis, sentence length turns out to be not feasible for discrimination. However, this results does not depreciate sentence length as an important stylistic factor; rather functional styles turn out to be a socio-linguistic rather than a stylistic category. The same conclusion has to be drawn with regard to text sorts.
b. with regard to our 333 Slovenian texts four statistical characteristics turn out to be relevant in discriminant analyses, based on sentence length as the only discriminating factor: mean sentence length ($\bar{x}$), standard deviation (s), Ord's S, and enropy h; at least these variables should be taken into account in future studies, though it may well turn out that other variables play a more decisive role;
c. our discriminant analyses result in a new text typology, involving six major text groups: in this typology, sentence length has a strong discriminating power particularly for dramatic texts (oral discourse), cooking recipes (technical discourse), and novel texts (everyday narration); with certain reservations, this holds true for scientific and journalistic discourse, too, with some transitions between these two discourse types. Only private letters represent a relatively heterogeneous group which cannot clearly be attributed to one of the major discourse types.

Given these findings, it will be tempting to compare the results obtained to those, previously gained on the basis of word length as discriminating variable. On the one hand, this will provide insight into the power of two (or more) combined linguistic variables for questions of text classification; it will be particularly interesting to see in how far classifications obtained on the basis of other variables (or specific combinations of variables) lead to identical or different results. Finally, insight will be gained into the stylistic structure of specific texts, and discourse types, in a more general understanding.

References

ADAMZIK, K. (Ed.) (2000): *Textsorten. Reflexionen und Analysen.* Stauffenburg, Tübingen.

ALEKSEEV, P.M. (1988): *Kvantitativnaja lingvistika teksta.* LGU, Leningrad.

ANTIĆ, G., KELIH, E.; GRZYBEK, P. (2005): Zero-syllable Words in Determining Word Length. In: P. Grzybek (Ed.): *Contributions to the science of language. Word Length Studies and Related Issues.* Kluwer, Dordrecht, 117–157.

COPECK, T., BARKER, K., DELISLE, S. and SZPAKOWICZ, St. (2000): Automating the Measurement of Linguistic Features to Help Classify Texts as Technical. In: *TALN-2000, Actes de la 7^{e} Conférence Annuelle sur le Traitement Automatique des Langues Naturelle, Lausanne, Oct. 2000, 101–110.*

GRZYBEK, P. (Ed.) (2005): *Contributions to the Science of Language. Word Length Studies and Related Issues.* Kluwer, Dordrecht.

GRZYBEK, P. and KELIH, E. (2005): Textforschung: Empirisch! In: J. Banke, A. Schröter and B. Dumont (Eds.): *Textsortenforschungen.* Leipzig. [In print]

GRZYBEK, P., STADLOBER, E., KELIH, E., and ANTIĆ, G. (2005): Quantitative Text Typology: The Impact of Word Length. In: C. Weihs and W. Gaul (Eds.), *Classification – The Ubiquitous Challenge.* Springer, Heidelberg; 53–64.

KARLGREN, J. and CUTTING, D. (1994): Recognizing text genres with simple metrics using discriminant analysis. In: M. Nagao (Ed.): *Proceedings of COLING 94, 1071–1075.*

KELIH, E., ANTIĆ, G., GRZYBEK, P. and STADLOBER, E. (2005) Classification of Author and/or Genre? The Impact of Word Length. In: C. Weihs and W. Gaul (Eds.), *Classification – The Ubiquitous Challenge.* Springer, Heidelberg; 498–505.

KELIH, E. and GRZYBEK, P. (2005): Satzlängen: Definitionen, Häufigkeiten, Modelle. In: A. Mehler (Ed.), *Quantitative Methoden in Computerlinguistik und Sprachtechnologie. [= Special Issue of: LDV-Forum. Zeitschrift für Computerlinguistik und Sprachtechnologie / Journal for Computational Linguistics and Language Technology]* [In print]

OHNHEISER, I. (1999): Funktionale Stilistik. In: H. Jachnow (Ed.): *Handbuch der sprachwissenschaftlichen Russistik und ihrer Grenzdisziplinen.* Harrassowitz, Wiesbaden, 660–686.

SMITH, M.W.A (1983): Recent Experience and New Developments of Methods for the Determination of Authorship. *Bulletin of the Association for Literary and Linguistic Computing, 11(3), 73–82.*