# Quantitative Text Typology: The Impact of Word Length

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

[1] Department for Slavic Studies, University Graz, A-8010 Graz, Austria
[2] Department for Statistics, Graz University of Technology, A-8010 Graz, Austria

**Abstract.** The present study aims at the quantitative classification of texts and text types. By way of a case study, 398 Slovenian texts from different genres and authors are analyzed as to their word length. It is shown that word length is an important factor in the synergetic self-regulation of texts and text types, and that word length may significantly contribute to a new typology of discourse types.[1]

## 1 Introduction: Structuring the Universe of Texts

Theoretically speaking, we assume that there is a universe of texts representing an open (or closed) system, i.e. an infinite (or finite) number of textual objects. The structure of this universe can be described by two processes: identification of its objects, based on a definition of 'text', and classification of these objects, resulting in the identification and description of hierarchically ordered sub-systems. To pursue the astronomic metaphor, the textual universe will be divided into particular galaxies, serving as attractors of individual objects. Finally, within such galaxies, particular sub-systems of lower levels will be identified, comparable to, e.g., stellar or solar systems.

The two processes of identification and classification cannot be realized without recourse to theoretical assumptions as to the obligatory and/or facultative characteristics of the objects under study: neither quantitative nor qualitative characteristics are immanent to the objects; rather, they are the result of analytical cognitive processes.

### 1.1 Classification and Quantification

To one degree or another, any kind of classification involves quantification: Even in seemingly qualitative approaches, quantitative arguments come into play, albeit possibly only claiming – implicitly or explicitly – that some objects are 'more' or 'less' similar or close to each other, or to some alleged norm or prototype. The degree of quantification is governed by the traits incorporated into the meta-language. Hence it is of relevance on which analytical

[1] This study has been conducted in context of research project # 15485 («Word Length Frequencies in Slavic Texts»), financially supported by the Austrian Research Fund (FWF).

level the process of classification is started. Note that each level has its own problems as to the definition of sub-systems and their boundaries.

In any case, a classification of the textual universe cannot be achieved without empirical research. Here, it is important to note that the understanding of empirical work is quite different in different disciplines, be they concerned with linguistic objects or not. Also, the proportion of theory and practice, the weighting of qualitative and quantitative arguments, may significantly differ. Disciplines concentrating on language tend to favor theoretical and qualitative approaches. Aside from these "traditional" approaches, corpus linguistics as a specific sub-discipline dealing with language(s) and texts, has such a predominant empirical component. Defining itself as a "data-oriented" discipline, the basic assumption of corpus linguistics is that a maximization of the data basis will result in an increasingly appropriate ("representative") language description. Ultimately, none of these disciplines – be they of predominantly theoretical or empirical orientation – can work without quantitative methods.

Here, quantitative linguistics comes into play as an important discipline in its own right: as opposed to the approaches described above, quantitative linguistics strives for the detection of regularities and connections in the language system, aiming at an empirically based theory of language. The transformation of observed linguistic data into quantities (i.e., variables and constants), is understood as a standardized approach to observation. Specific hypotheses are statistically tested and, ideally speaking, the final interpretation of the results obtained is integrated into a theoretical framework.

### 1.2 Quantitative Text Analysis: From a Definition of the Basics Towards Data Homogeneity

The present attempt follows these lines, striving for a quantitative text typology. As compared to corpus linguistics, this approach – which may be termed *quantitative text analysis* – is characterized by two major lines of thinking: apart from the predominantly theoretical orientation, the assumption of quantitative text analysis is that 'text' is the relevant analytical unit at the basis of the present analysis.

Since corpus linguistics aims at the construction, or re-construction, of particular norms, of "representative" standards, of (a given) language, corpus-oriented analyses are usually based on a mixture of heterogeneous texts, of a "quasi text", in a way (Orlov 1982). On contrast, quantitative text analysis focuses on texts as homogeneous entities. The basic assumption is that a (complete) text is a self-regulating system, ruled by particular regularities. These regularities need not necessarily be present in text segments, and they are likely to intermingle in any kind of text combination. Quite logically, the question remains, what a 'text' is: is it a complete novel, composed of books?, or the complete book of a novel, consisting of several chapters?, or each individual chapter of a given book?, or perhaps even a paragraph, or a dialogical

or narrative sequence within it? Ultimately, there is no clear definition in text scholarship, and questions whether we need a "new" definition of text, regularly re-occur in relevant discussions. Of course, this theoretical question goes beyond the scope of this paper. From a statistical point of view, we are concerned with two major problems: the problem of data homogeneity, and the problem of the basic analytical units. Thus, particular decisions have to be made as to the boundary conditions of our study:

- ▷ We consider a *'text'* to be the result of a homogeneous process of text generation. Therefore, we concentrate on letters, or newspaper comments, or on chapters of novels, as individual texts. Assuming that such a 'text' is governed by synergetic processes, these processes can and must be quantitatively described. The descriptive models obtained for each 'text' can be compared to each other, possibly resulting in one or more general model(s); thus, a quantitative typology of texts can be obtained.
- ▷ But even with a particular definition of 'text', it has to be decided which of their traits are to be submitted to quantitative analyses. Here, we concentrate on *word length*, as one particular linguistic trait of a text.

### 1.3 Word Length in a Synergetic Context

Word length is, of course, only one linguistic trait of texts, among others, and one would not expect a coherent text typology, based on word length only. However, the criterion of word length is not an arbitrarily chosen factor (cf. Grzybek 2004). First, experience has shown that genre is a crucial factor influencing word length (Grzybek/Kelih 2004; Kelih et al., this volume); this observation may as well turned into the question to what degree word length studies may contribute to a quantitative typology of texts. And second, word length is an important factor in a synergetic approach to language and text. We cannot discuss the synergetics of language in detail, here (cf. Köhler 1986); yet, it should be made clear that word length is no isolated linguistic phenomenon: given one accepts the distinction of linguistic levels, as (1) phoneme/grapheme, (2) syllable/morpheme, (3) word/lexeme, (4) clause, and (5) sentence, at least the first three levels are concerned with recurrent units. Consequently, on each of these levels, the re-occurrence of units results in particular frequencies, which may be modelled with recourse to specific frequency distribution models. Both the units and their frequencies are closely related to each other. The units of all five levels are characterized by length, again mutually influencing each other, resulting in specific frequency length distributions. Table 1 demonstrates the interrelations.

Finally, in addition to the decisions made, it remains to be decided which shall be the analytical units, that is not only what a 'word' is (a graphemic, phonetic, phonological, intonational, etc. unit), but also in which units word length is supposed to be measured (number of letters, of graphemes, of phonemes, syllables, morphemes, etc.).

**Table 1.** Word Length in a Synergetic Circuit

| | | | |
|---|---|---|---|
| | SENTENCE | Length | Frequency |
| | | ↕ | |
| | CLAUSE | Length | Frequency |
| | ↱ | ↰ ↕ | |
| Frequency | WORD / LEXEME | Length | Frequency |
| ↕ | ↱ | ↰ ↕ | |
| Frequency | SYLLABLE / MORPHEME | Length | Frequency |
| ↕ | ↱ | ↰ ↕ | |
| Frequency | PHONEME / GRAPHEME | Length | Frequency |

▷ In the present analysis, we concentrate on word as an orthographic-phonemic category (cf. Antić et al. 2004), measuring word length as the number of syllables per word.

### 1.4 Qualitative and Quantitative Classifications: A Priori and A Posteriori

Given these definitions, we can now pursue our basic question as to a quantitative text typology. As mentioned above, the quantitative aspect of classification is often neglected or even ignored in qualitative approaches. As opposed to this, qualitative categories play an overtly accepted role in quantitative approaches, though the direction of analysis may be different:

1. One may favor a *"tabula rasa" principle* not attributing any qualitative characteristics in advance; the universe of texts is structured according to word length only, e.g. by clustering methods, by analyzing the parameters of frequency distributions, etc.;
2. One may prefer an *a priori* ↔ *a posteriori principle*: in this case, a particular qualitative characteristic is attributed to each text, and then, e.g. by discriminant analysis, one tests whether these categorizations correspond to the quantitative results obtained.

Applying qualitative categories, the problem of data heterogeneity once again comes into play, now depending on the meta-language chosen. In order to understand the problem, let us suppose, we want to attribute a category such as 'text type' to each text. In any kind of qualitative approaches, the text universe is structured with regard to external (pragmatic) factors – "with reference to the world", in other words. Either, these categories are based on general communicative functions of language, which results in the distinction of particular functional styles; or, they are based on specific situational functions, thus resulting in the distinction of specific text sorts.

(a) One procedure is to classify each text according to the functional style it belongs to. The concept of functional style, successfully applied in previous quantitative research (cf. Mistrík 1966), has been mainly developed in Russian and Czechoslovak stylistics, where style has been understood as

serving particular socio-communicative functions. A functional style thus relates to particular discourse spheres: everyday, official-administrative, scientific, journalistic, or artistic styles. Such a coarse categorization with about half a dozen of categories, may result in an enormous heterogeneity of the individual texts included in the individual categories.

(b) As opposed to this, relying on contemporary studies of so-called text sorts (cf. Adamczik 1995: 255ff.), there are more than 4,000 categories at our disposal. In this case, our categories are less broad and general, the material included tends to be more homogeneous, but the number of categories can hardly be handled in empirical research.

In order to profit from the advantages of both approaches, it seems reasonable to combine these two principles. In this combination (cf. Grzybek/Kelih 2004), each text sort is attributed to one of the functional styles (cf. Figure 1). In a quantitative approach, such an attribution can only be understood as an a priori classification. Anyway, both bottom-up (text → text sort → functional style) and top-down analyses are possible in a vertical perspective, as well as first order and second order cross-comparisons, in a horizontal perspective (i.e., between different functional styles or text sorts).

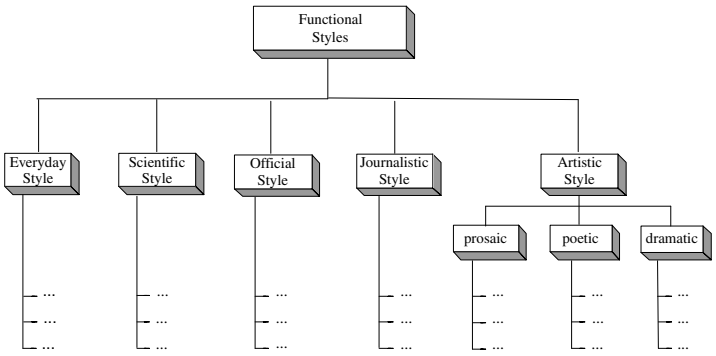



**Fig. 1.** Functional Styles and Text Sorts

Our basic assumption is that the highest level – the entities of which are comparable to 'text galaxies' (see above) – should not primarily considered to be defined by socio-communicative functions, but regarded as linguistic phenomena: It seems reasonable to assume that different text sorts (analogous to our "stellar systems"), which serve particular functions as well, should be characterized by similar linguistic or stylistic traits. As opposed to merely qualitative text typologies, the attribution of text sorts to functional styles is to be understood as an a priori hypothesis, to be submitted to empirical tests. As a result, it is likely that either the a priori attributions have to be modified, or that other categories have to be defined at the top level, e.g. specific *discourse types*, instead of functional styles.

## 2 A Case Study: Classifying 398 Slovenian Texts

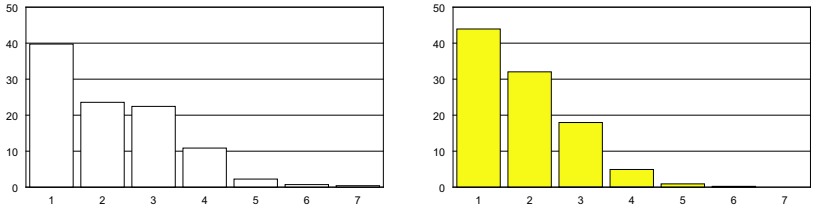
The present case study is an attempt to arrive at a classification of 398 Slovenian texts, belonging to various sorts, largely representing the spectrum of functional styles; the sample is represented in Table 2.

**Table 2.** 398 Slovenian Texts

| FUNCTIONAL STYLE | AUTHOR(S) | TEXT TYPE(S) | no. |
|---|---|---|---|
| ⊡ **Everyday** | Cankar, Jurčič | Private Letters | 61 |
| ⊡ **Public** | various | Open Letters | 29 |
| ⊡ **Journalistic** | various | Readers' Letters, Comments | 65 |
| ⊡ **Artistic** | | | |
| ⊙ *Prose* | Cankar | Individual Chapters from Short Novels (*povest*) | 68 |
| | Švigelj-Mérat / Kolšek | Letters from an Epistolary Novel | 93 |
| ⊙ *Poetry* | Gregorčič | Versified Poems | 40 |
| ⊙ *Drama* | Jančar | Individual Acts from Dramas | 42 |

The obvious emphasis on different types of letters in the sample is motivated by the fact that 'letter' as a genre often is regarded to be prototypical of language in general: thus, 'letter' as a genre is assumed to be located between oral and written communication, and it is considered to be the result of a unified, homogeneous process of text generation. This assumption turns out to be problematic, however, if one takes into account the fact that in contemporary text sort research (cf. Adamczik 1995: 255ff.), several dozens of different letter types are distinguished. Consequently, it would be of utmost importance (a) to compare how the genre of letters as a whole relates to other genres, and (b) to see how different letter types relate to each other – in fact, any difference between different letter types would weaken the argument in favor of 'letter' being a homogeneous prototype of language.

In our analyses, each text is analyzed with regard to word length. Average word length ($m_1$) is only one of many possible variables, of course, which may characterize a given frequency distribution. In fact, a pool of variables has been developed in the project mentioned above (cf. fn. 1), consisting of 30 variables, such as the four central moments, variance and standard deviation, coefficient of variation, dispersion index, entropy, repeat rate, etc. All these variables are derived from the individual word length frequencies of a given text; by way of an example, Figure 2 represents the relative frequencies of $x$-syllable words for two arbitrarily chosen texts. In this case, there are significant differences between almost all length classes.




(a) Journalistic Comment (#324) (b) Private Letter (#1)

**Fig. 2.** Word Length Frequencies (in %) of Two Different Texts

### 2.1 Post Hoc Analysis of Mean Word Length

By way of a first approximation, it seems reasonable to calculate a post-hoc-analysis of the mean values. As a result of this procedure, groups without significant differences form homogeneous subgroups, whereas differing groups are placed in different groups. As can be seen from Table 3, which is based on mean word length ($m_1$) only, homogeneous subgroups do in fact exist; even more importantly, however, all four letter types fall into different categories. This sheds serious doubt on the assumption, that 'letter' as a category might serve as a prototype of language without further distinction.

**Table 3.** Post Hoc Analyses ($m_1$)

| | | Subgroup for $\alpha = .05$ | | | | |
|---|---|---|---|---|---|---|
| Text sort | $n$ | 1 | 2 | 3 | 4 | 5 |
| Poems | 40 | 1.7127 | | | | |
| Short stories | 68 | | 1.8258 | | | |
| Private letters | 61 | | 1.8798 | | | |
| Drama | 42 | | 1.8973 | | | |
| Epistolary novel | 93 | | | 2.0026 | | |
| Readers' letters | 30 | | | | 2.2622 | |
| Comments | 35 | | | | 2.2883 | |
| Open Letters | 29 | | | | | 2.4268 |

### 2.2 Discriminant Analyses: The Whole Corpus

In linear discriminant analyses, specific variables are submitted to linear transformations in order to arrive at an optimal discrimination of the individual cases. At first glance, many variables of our pool may be important for discrimination, where the individual texts are attributed to groups, on the basis of these variables. However, most of the variables are redundant due to

their correlation structure. The stepwise procedures in our analyses resulted in at most four relevant predictor variables for the discriminant functions. Figure 3 shows the results of the discriminant analysis for all eight text sorts, based on four variables: mean word length ($m_1$), variance ($m_2$), coefficient of variation ($v = s/m_1$), and relative frequency of one-syllable words ($p_1$). With only 56.30% of all texts being correctly discriminated, some general tendencies can be observed: (1) although some text sorts are located in clearly defined areas, there are many overlappings; (2) poems seem to be a separate category, as well as readers' letters, open letters, and comments, on the other end; (3) drama, short story, private letters and the letters from the epistolary novel seem to represent some vaguely defined common area.

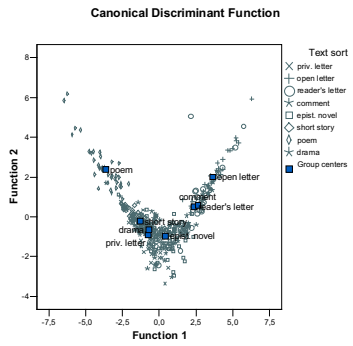



**Fig. 3.** Discriminant Analysis: Eight Text Sorts

### 2.3 From Four to Two Letter Types

In a first approach to understand the underlying structure of the textual universe, let us concentrate on our four letter types, only, since they were all attributed to different classes in the post hoc analyses. Treating all of them – i.e., private letters ($PL$), open letters ($OL$), readers' letters ($RL$), and letters from an epistolary novel ($EN$) –, as a separate class, one arrives at a percentage of 70.40% correctly discriminated texts. This result is obtained with two relevant variables: $m_1$ and $v$.

**Table 4.** Discriminant Analysis: Four Letter Types ($n = 213$)

| | Predicted group | | | | |
|---|---|---|---|---|---|
| Letter Type | $PL$ | $OL$ | $RL$ | $EN$ | Total |
| $PL$ | 37 | 0 | 2 | 22 | 61 |
| $OL$ | 0 | 22 | 3 | 4 | 29 |
| $RL$ | 1 | 9 | 10 | 10 | 30 |
| $EN$ | 10 | 0 | 3 | 80 | 93 |

There is an obvious tendency that private letters ($PL$) and the letters from the epistolary novel ($EN$) represent a common category, whereas open letters ($OL$) and readers's letters ($RL$) display this tendency to a lesser degree, if at all. In fact, combining private letters and the letters from the epistolary novel

in one group, thus discriminating three classes of letters, yields a percentage of 86.90% correctly discriminated texts, with only two variables: $m_1$ and $p_2$ (i.e., the percentage of two-syllable words). Table 5 shows the results in detail.

**Table 5.** Discriminant Analysis: Three Letter Types ($n = 213$)

| | Predicted group | | | |
|---|---|---|---|---|
| Group | 1 | 2 | 3 | Total |
| 1 | 151 | 0 | 3 | 154 |
| 2 | 2 | 20 | 6 | 28 |
| 3 | 12 | 5 | 14 | 31 |

1=$\{PL, EN\}$ 2=$OL$ 3=$RL$

As Table 5 shows, 98% of the combined group are correctly discriminated. This is a strong argument in favor of the assumption that we are concerned with some common group of private letters, be they literary or not. This result sheds serious doubt on the possibility to distinguish private letters from literary letters, at least from a quantitative point of view. Obviously, the literary letters reproduce or "imitate" the linguistic style of private letters, what calls into question the existence of the functional style of prosaic literature. Given this observation, it seems reasonable to combine readers' letters ($RL$) and open letters ($OL$) in one common group, too, and to juxtapose this group of public letters to the group of private letters. In fact, this results in a high percentage of 92.00%, with $m_1$ and $p_2$ being the relevant variables.

### 2.4 Towards a New Typology

On the basis of these findings, the question arises if the two major groups – private letters ($PL/EN$) and public letters ($OL/RL$) – are a special case of more general categories, such as, e.g., 'private/everyday style' and 'public/official style'. If this assumption should be confirmed, the re-introduction of previously eliminated text sorts should yield positive results.

**The re-introduction of journalistic comments** ($CO$) to the group of public texts does not, in fact, result in a decrease of the good discrimination result: as can be seen from Table 6, 91.10% of the 248 texts are correctly discriminated (again, with $m_1$ and $p_2$, only). Obviously, some distinction along the line of public/official vs. private/everyday texts seems to be relevant.

**The re-introduction of the dramatic texts** ($DR$), as well, seems to be a logical consequence, regarding them as the literary pendant of everyday dialogue. We thus have 290 texts, originating from six different text sorts,

**Table 6.** Discriminant Analysis: Five Text Sorts in Two Categories: Public/Official vs. Private/Everyday ($n = 248$)

| | Predicted group | | |
|---|---|---|---|
| Group | 1 | 2 | Total |
| 1 | 148 | 6 | 154 |
| 2 | 16 | 78 | 94 |
| 1=$\{PL, EN\}$ | 2=$\{OL, RL, CO\}$ | | |

and grouped in two major classes; in fact, as can be seen from Table 7; 92.40% of the texts are correctly discriminated. One might object, now, that the consideration of only two classes is likely to be effective. Yet, it is a remarkable result that the addition of two non-letter text sorts does not result in a decrease of the previous result.

**Table 7.** Discriminant Analysis: Six Text Sorts in Two Categories: Public/Official vs. Private/Everyday ($n = 290$)

| | Predicted group | | |
|---|---|---|---|
| Group | 1 | 2 | Total |
| 1 | 190 | 6 | 196 |
| 2 | 16 | 78 | 94 |
| 1=$\{PL, EN, DR\}$ | 2=$\{OL, RL, CO\}$ | | |

**The re-introduction of the poetic texts** ($PO$) as a category in its own right, results in three text classes. Interestingly enough, under these circumstances, too, the result is not worse: rather, a percentage of 91.20% correct discriminations is obtained on the basis of only three variables: $m_1, p_2, v$. The results are represented in detail, in Table 8.

**Table 8.** Discriminant Analysis: Seven Text Sorts in Three Categories: Public/Official vs. Private/Everyday vs. Poetry ($n = 330$)

| | Predicted group | | | |
|---|---|---|---|---|
| Group | 1 | 2 | 3 | Total |
| 1 | 191 | 3 | 2 | 196 |
| 2 | 19 | 75 | 0 | 94 |
| 3 | 5 | 0 | 35 | 40 |
| 1=$\{PL, EN, DR\}$ | 2=$\{OL, RL, CO\}$ | 3=$\{PO\}$ | | |

Figure 4 illustrates the result of the discriminant analysis. It can clearly be seen that the poetic texts and the public texts are located at the two extremes of the textual spectrum, and that they do not represent any case of the overall few mis-classifications. At this point, the obvious question arises if a new typology might be the result of our quantitative classification. With this perspective in mind, it should be noticed that seven of our eight text sorts are analyzed in Table 8.

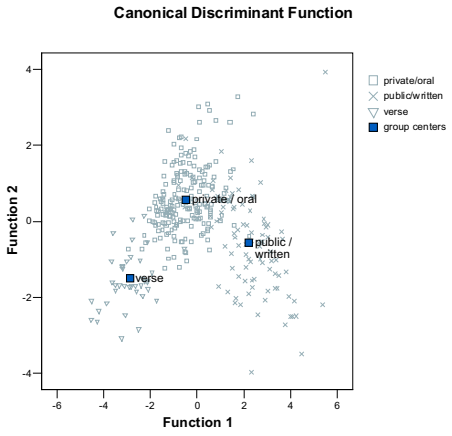



**Fig. 4.** Discriminant Analysis: Three Discourse Types

**The re-introduction of the literary prose texts** ($LP$) is the last step, thus again arriving at the initial number of eight text sorts. As can be seen from Table 9, the percentage of correctly discriminated texts now decreases to 79.90%.

**Table 9.** Discriminant Analysis: Eight Text Sorts in Four Categories ($n = 398$)

| | Predicted group | | | | |
|---|---|---|---|---|---|
| Group | 1 | 2 | 3 | 4 | Total |
| 1 | 183 | 3 | 9 | 1 | 196 |
| 2 | 19 | 75 | 0 | 0 | 94 |
| 3 | 42 | 0 | 26 | 0 | 68 |
| 4 | 1 | 0 | 5 | 34 | 40 |

1=$\{PL, EN, DR\}$ 2=$\{OL, RL, CO\}$
3=$\{LP\}$ 4=$\{PO\}$

A closer analysis shows that the most mis-classifications appear between literary texts and private letters. Interestingly enough, many of these texts are from one and the same author (Ivan Cankar). One might therefore suspect authorship to be an important factor; however, Kelih et al. (this volume) have good arguments (and convincing empirical evidence) that word length is less dependent on authorship, than it is on genre. As an alternative interpretation, the reason may well be a specific for the analyzed material because in case of the literary texts, we are concerned with short stories which aim at the imitation of orality, and include dialogues to varying degree.

Therefore, including the literary prose texts ($LP$) in the group of inofficial/oral texts, and separating them from the official/written group, on the one hand, and the poetry group, on the other, results in a percentage of 92.70% correctly discriminated texts, as can be seen from Table 10.

**Table 10.** Discriminant Analysis: Eight Text Sorts in Three Categories: Inofficial / Oral vs. Official / Written vs. Poetry ($n = 398$)

| | Predicted group | | | |
|---|---|---|---|---|
| Group | 1 | 2 | 3 | Total |
| 1 | 260 | 3 | 1 | 264 |
| 2 | 19 | 75 | 0 | 94 |
| 3 | 6 | 0 | 34 | 40 |

1=$\{PL, EN, DR, LP\}$ 2=$\{OL, RL, CO\}$ 3=$\{PO\}$

The final outcome of our classification is represented in Figure 5.

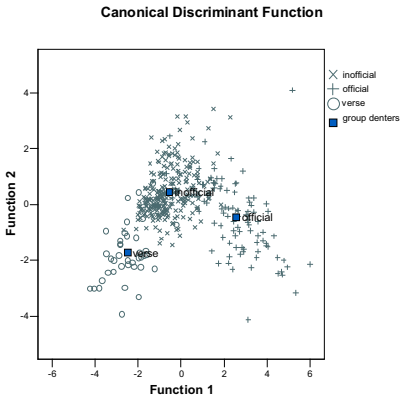



**Fig. 5.** (Discriminant Analysis: Final Results and New Categorization)

The results suggest there to be (at least) three different discourse types, which do not coincide with the traditional distinction of functional styles. It remains to be seen in future, what kind of additional discourse types will have to be distinguished, and how further text sorts relate to this scheme.

## References

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

ANTIĆ, G., KELIH, E., and GRZYBEK, P. (2004): Zero-syllable Words in Determining Word Length. In: P. Grzybek ( Ed.): *Contributions to the Science of Language. Word Length Studies and Related Issues.* [In print]

GRZYBEK, P. (2004): History and Methodology of Word Length Studies: The State of the Art. In: P. Grzybek (Ed.): *Contributions to the Science of Language: Word Length Studies and Related Issues.* [In print]

GRZYBEK, P. and KELIH, E. (2004): Texttypologie in/aus empirischer Sicht. In: J. Bernard, P. Grzybek, and Ju. Fikfak (Eds.): *Text and Reality.* Ljubljana etc. [In print].

GRZYBEK, P. and STADLOBER, E. (2003): Zur Prosa Karel Čapeks – Einige quantitative Bemerkungen. In: S. Kempgen, U. Schweier, and und T. Berger (Eds.), *Rusistika – Slavistika – Lingvistika. Festschrift für Werner Lehfeldt zum 60. Geburtstag.* Sagner, München, 474–488.

KELIH, E., ANTIĆ, G., GRZYBEK, P., and STADLOBER, E. (2004): Classification of Author and/or Genre? [Cf. this volume]

KÖHLER, R. (1986): Zur synergetischen Linguistik: Struktur und Dynamik der Lexik. Brockmeyer, Bochum.

ORLOV, Ju.K. (19): Linguostatistik: Aufstellung von Sprachnormen oder Analyse des Redeprozesses? (Die Antinomie «Sprache–Rede» in der statistischen Linguistik). In: Ju.K. Orlov; M.G. Boroda, I.Š. Nadarešvili: *Sprache, Text, Kunst. Quantitative Analysen.* Brockmeyer, Bochum.