On functional and computational LSP analysis

Renkema, J.

Published in:
Reading for professional purposes

Publication date:
1984

10 On functional and computational LSP analysis: the example of officialese

J. Renkema

Introduction

Research on language for special purposes has become easier in recent years thanks to progress in computational linguistics. In this article I would like to show some results of computational research on LSP. But first of all some statements may be made about useful research in this field.

1. The description of an LSP must be related to the function a special kind of language has. For example, one can study the frequency of plural and singular nouns in newspaper language, but it is not reasonable to expect that results of such study will give insight into the characteristics of newspaper language. Of course, one can count every surface phenomenon, but mere counting, without relating results to the function of the type of language under consideration, would have little to do with linguistic research.

2. The description of an LSP can only be interesting for our purposes when it reveals the differences between that language and other, related LSPs. It is not useful for our purposes to describe differences between, for example, poetical language and the language used in cookery books. We should, however, gain some insight into the domain of LSP when we succeed in describing differences between, for example, the language used in cookery books and the language of 'directions for use'. This means in fact that LSP-description has much to do with a stylistic description within the framework of style as a deviation from a *related* kind of language use. (For a detailed discussion see Renkema, 1981.)

3. Although the use of a computer makes it possible to study large samples of LSP, it should be emphasized that only very clearly defined phenomena can be objects of research. Ill-defined lexical characteristics, such as the occurrence of technical terms, can only be studied when there is a clear definition of what is meant by technical terms. Of course, clear definitions are also necessary in research without computers, but a machine that can only follow instructions in terms of ordering and counting, compels the researcher to provide definitions without any ambiguous element.

4. When one studies an LSP it is not useful - at least in my opinion - to study *only* specific characteristics related to the particular functions of one LSP. As noted above, a complementary approach is necessary in order to relate different LSPs and to gain insight into the possibilities of language variation. Therefore general parameters should be used wherever possible to

place the LSP under consideration in the whole range of language varieties.

So much for introductory statements: in what follows I shall try to illustrate these statements with my study of Dutch officialese. First I shall make some remarks on the function of officialese (1). Then I shall give some data about the kinds of language use which are studied in this framework (2). After that I shall give a short glimpse of computer research into this type of language use and present results concerning some lexical parameters (3). Finally I shall make some remarks on the general parameter type-token ratio (TTR), and present the first results of research on TTR by word category (4).

1. The LSP officialese

What is the function of officialese? Before I try to formulate an answer, some distinctions are necessary. In my opinion, the term officialese can refer to three different types of language: 1 legal language, 2 political language, 3 bureaucratic language. It is impossible to make clear distinctions between these types of language use, but there are some differences in function.

The dominant function of legal language is to formulate judicial regulations in such a way that they are unambiguous. The communication function of legal language is diminished when someone finds loopholes in the law. So, legal language has to be very precise, with many additions, exceptions, and so on. Here, of course, 'precise' is not the same as 'easy to read'.

Political language, however, has quite different functions, which could be described – with some simplification – as 'trying to win an election' on the one hand, and formulating compromises on the other hand. So political language has to be rather persuasive, wordy and diplomatic.

A dominant function of bureaucratic language – in my view – is to be informative to many readers in explaining legal points to citizens and in providing substantive resources for political discussion. So this kind of language use has to be *factual*. The purpose of this language is to elicit action, so it has to be rather *terse* and concise without drawing attention to the language itself. Furthermore, the legal points and the data for political discussion are not products of personal feelings, but have to be seen as the output of an official body (rather than a group of people). Hence the language has to be *objective*. A fourth characteristic worthy of note is that bureaucratic actions are produced in the name of High Authority. So the language has to be somewhat *dignified* or stately. Some of the functions mentioned could also apply to legal or even political language. But in my view they are not dominant in these kinds of language use.

In this study I am concerned with bureaucratic language. Legal and political language have dominant functions so different that they cannot profitably be studied in the same way as bureaucratic language. I suppose that much discussion is possible about these four characteristics of bureaucratic language: 1 matter-of-fact-ness, 2 terseness, 3 objectivity and 4 dignity. I can refer to many critics of this type of language use in different

language areas who in fact mention the same characteristics, but in negative terms, for example bureaucratic language is: dry as dust, complicated, impersonal and traditional. (See Renkema 1977, 1981). But even if one does not fully agree with these rough characteristics, I still want to stick to my point that if an LSP is studied, one has to take as a starting point the functions the LSP in question is, rightly or wrongly, supposed to have. Only in this way is there a guarantee that the right kind of phenomena are studied. Of course the emphasis on functions as a starting point does not mean that every formal characteristic can be explained by the function it has. The form-function relation, however, is too intricate to be discussed here.

2. On computational data of language varieties

Before I give an illustration of what I have studied within the framework of my functional description of bureaucratic language, I shall give some information on procedures developed to analyse the language varieties which I consider to be types of language use nearest in function, namely: the language in daily and weekly newspapers, and the language in popular scientific prose. I consider these types of language use to be related varieties because of their similarity in informative function and their purposes, namely to reach many readers. Large samples of these varieties (120,000 words per category) had already been encoded in a comprehensive encoding system of grammatical and lexical description, the C3C system (see Uit den Boogaart, 1975). For every word in a text this system provides a number code of three digits which (420) specifies (253) its (330) grammatical (100) function (000) and (700) relevant (100) characteristics (001) of (600) form (000).[1]

For my research the encoded samples were stored in a computer for analysis by means of the 'Eye on Text' program. This program was especially designed for analysing texts at the University of Amsterdam, Department for General Linguistics and Computer Linguistics (Brandt Corstius, 1978).

The program enables the computer to produce from a corpus any sentences that meet certain specified conditions. A condition can be formulated as a string of words and codes, or both, connected by the logical operators *and, or* and *not*. There are three possible variables: (*) a word, (c) a code, (-) a string of words with codes. It is also possible to state that a condition has to occur at the beginning of the sentence or, for example, at the seventh word. For example, when someone is looking for data on participles used substantively - at least in Dutch - the inflected and plural participles have to be considered. The C3C system contains special codes for these categories. But not all forms in these categories are used substantively.

[1]An example: 'specifies 253' means: 'specifies' is a verb (2) used transitively (5) as finite form in the third person singular (3).

Compare *het volgende* 203 (the following) versus *de volgende* 203 *opmerkingen* 001[2] (the following remarks). Via 'Eye on Text' it is possible to call up sentences with the condition (*203 - and not (*203. and *0 . .)). With this simple operation it is possible to select from a large sample all sentences with pure substantively used participles. This example shows that one can answer very detailed questions on differences in vocabulary and syntax with the program 'Eye on Text' (For detailed information see Renkema, 1981).

Thus we have a possible analysis for language varieties which can be considered related types of language use. For my research on bureaucratic language I have selected a sample of government publications. Random fragments (totalling 48,242 words) were taken from the correspondence between the government and the States General (the Dutch Parliament) for the parliamentary year 1975-1976. This sample is encoded in the C3C encoding system and then stored in a computer for analysis via 'Eye to Text'. Although subject to justified criticism, the C3C system is used here because it is the only system which has actually been applied on a large scale.

3. Some data on bureaucratic language

I will now give some results to illustrate the way this LSP can be studied. One of the functions of bureaucratic language I mentioned was matter-of-factness. With a view to studying this function I decided to explore the number of nouns.

If there is more matter-of-factness in bureaucratic language than in newspaper language or popular scientific prose, then it is reasonable to hypothesize that bureaucratic language contains more nouns, or other words used as nouns, and fewer verbs; i.e. if there is more matter-of-factness then there will be more 'nominal style'. If there are more nouns than verbs, then maybe some nouns are derived from verbs, e.g. infinitives and participles used as nouns. In Dutch there is a very productive suffix for making nouns from verbs, namely *-ing*, like in English run (verb) and the running (noun). In Table 1 I have mentioned some hypotheses concerning nominal style and the results.

In this table the hypotheses are mentioned in headwords. (+) or (-) indicate that the linguistic phenomenon in question has a higher or lower frequency in bureaucratic language than in dailies, weeklies or in popular scientific prose.

So I found that bureaucratic language contains significantly ($p < 0.001$) more nouns (including nouns ending in *-ing*) than the language of newspapers and popular scientific prose. But there are no significant differences with respect to participles used as nouns. Furthermore the

[2]The code 203 means: a verb (2) used intransitively (0) as inflected present participle (3). Code 001 means: a noun (0) in ordinary way (0) - not a proper name etc. - in plural form (1). Code 0 . . means: a noun in general, including proper names etc., in all possible forms (singular, genitive, etc.).

Table 1
Nominal style in bureaucratic language

	D	W	PS
1. Nouns (+)	!	!	!
2. Infinitives used as nouns (+)	!	!	?
3. Participles used as nouns (+)	?	?	?
4. Verbs (-)	!	!	?
5. Nouns ending in '-ing' (+)	!	!	!

Symbols and abbreviations in Table 1
D Daily newspapers
W Weekly newspapers
PS Popular scientific prose
! Significant difference in expected direction
? Difference in expected direction, but not significant

popular scientific prose tends to be less different from bureaucratic than newspaper language, with respect to infinitives used as nouns and verbs.

This was only an illustration. There are more data available on matter-of-factness, and on objectivity, terseness and dignity (see Renkema, 1981). After this brief report of already published research I shall now turn to research in progress on a more general parameter - the type-token ratio (TTR) - in order to substantiate my fourth introductory statement that a study of one LSP can and should contribute to a general framework for language variety research.

4. The type-token ratio

The TTR is the quotient of the number of different words (types) and the total number of words (tokens) in a text. If in a text each word occurs once, then the TTR is 1. It will be clear that the TTR is always less than 1, because in a text many words occur more than once. The literature on TTR shows that many researchers have used this measure to produce relatively easily precise data on differences in vocabulary richness (see Renkema, forthcoming).

It is very doubtful if the type-token ratio is in fact, a reliable measure. Of course, the problem of validity is much more important than the problem of reliability: it is not clear at all what the TTR is supposed to measure. Given the state of the art, many investigations into stylostatistics have to be carried out within a psychological framework, as an attempt to solve the validity problem. But before we can deal with validity we must have a reliable TTR. Hence my emphasis here on reliability.

4.1. The reliability of the TTR

In my view there are three main problems:

1. *Text length*

It is clear that the TTR becomes lower as the text is longer, because a longer

text contains more tokens of frequent types (e.g. function words). So the TTR cannot be used for texts of different lengths. In the literature three solutions are offered. Herdan (1960) has developed a logarithmic TTR. But his proposals offer only a rough solution for samples with differences in length. Carroll (see e.g. Andolina, 1980) has suggested replacing the TTR by the quotient of types and $\sqrt{2x}$ tokens to neutralize differences in text length. But for me it is not clear why π or $\sqrt[3]{}$ or any other mathemetical adaptation should not fit here.

Manschreck *et al.* (1981) have worked with the mean segmental TTR (MS TTR). In this study TTRs are produced on the basis of the means of hundred word samples. I am – as a linguist – not capable of evaluating this approach. In my research I have worked with samples of exactly the same length and in reporting this work on TTR of word categories I shall give some information about what seems a more reliable measure for large samples.

2. *Homographs*

In TTR research no one, as far as I know, has faced the problem of homography. But for a precise TTR it is necessary to distinguish between the various meanings of a word. I give a Dutch example, the word *rond*. The meanings of this word include:

a. ze huppelen in het rond (000)[3]	they are hopping around
b. de bal is rond (100)	the ball is round
c. rond duizend gulden (500)	about a thousand guilders
d. hij hing maar wat rond (620)	he was just standing about

Thanks to the C3C encoding system it is possible to study texts in which these kinds of homographs are distinguished. In an encoded text one can count words with different code numbers. Then a type is redefined as a word with a different code instead of as a different word. However, homonyms or polysemes like 'table' (time table) and 'table' (for dinner) could not be distinguished. The word 'table' is always encoded as a noun without respect to its meaning. But the C3C encoding system offers at least a first step in distinguishing identical forms with different meanings.

3. *Characteristics of the sample*

The problem of sampling can be illustrated by the following example. In De Jong (1979) frequency tables are presented of varieties of spoken language in Dutch. With the four parameters – formality, sex, age and education, sixteen samples of 7,500 words each are encoded following the C3C encoding system. Note that in this way the problem of difference in text length is neutralized, and that problems of homographs are partly solved. In the following table I give the number of types and tokens and the TTR per variety.

[3]The codes of the C3C-system mean here:
000 substantive, ordinary – not proper name etc., basic form
100 adjective, ordinary – not used as adverb etc., basic form
500 adverb, ordinary – not interrogative etc., basic form
620 non-verbal part of a compound verb

Table 2
TTRs of spoken language varieties in Dutch

	Types	*Tokens*	*TTR*
1. F M Y H	1419	7,500	0.1892
2. F M Y L	1331	7,500	0.1774
3. F M O H	1604	7,500	0.2139
4. F M O L	1398	7,500	0.1864
5. F W Y H	1326	7,500	0.1768
6. F W Y L	1245	7,500	0.1666
7. F W O H	1383	7,500	0.1844
8. F W O L	1308	7,500	0.1744
9. I M Y H	1532	7,500	0.2042
10. I M Y L	1518	7,500	0.2043
11. I M O H	1565	7,500	0.2087
12. I M O L	1554	7,500	0.2072
13. I W Y H	1446	7,500	0.1928
14. I W Y L	1362	7,500	0.1816
15. I W O H	1400	7,500	0.1867
16. I W O L	1402	7,500	0.1869

F = formal
I = informal

O = old
Y = young

M = men
W = women

H = high education
L = low education

Statistical research by means of analysis of variance showed significant differences at the 5 per cent level on the four main factors:

- the TTR in formal language is lower than in informal language;
- men have a higher TTR than women;
- old people have a higher TTR than young people;
- people with a high education showed a higher TTR than people with a low education.

The existence of significant interactions of two or more factors could not be shown, though they may well be present.

It would be interesting to compare these results with intuitions about the relation between TTR and the four parameters. But instead of doing this I would like to query one of these results. When we look at the samples on which these results are based, it becomes clear that the findings as to the relation between TTR and degree of formality are not reliable as yet.

The corpus of formal spoken language contains interviews on ten different topics. But there are no data available about the number of topics in the

corpus of informal spoken language. It is reasonable to assume that the number of different topics can influence the TTR. Only after a precise survey of the topics in the 'informal' corpus can we make up our minds as to the reliability of these data. Therefore time-absorbing study of the samples is necessary. Thus we have seen one feature of the influence of the sample on TTR.

4.2. TTR of word categories

In this last section I want to give a survey of TTR research into word categories. Amazingly, in my opinion, up till now only overall TTRs have been studied. But I think it is more useful to study the TTR of different word categories because interesting problems in LSP research are often related to parts of speech, e.g. hypotheses concerning the richness of nouns in scientific articles or the variety of adjectives in advertisements.

For this study I have selected encoded samples of the same size as the sample of bureaucratic language (48, 242) not only from daily and weekly newspapers, and popular scientific prose, but also from novels and women's weeklies. The samples are stored in an Exidy Sorcerer microcomputer, with a RAM of 55 Kbyte. Also available were a matrix printer and a disk-drive for 8" diskettes. Of course the memory capacity is not sufficient for the large number of data. It was necessary to split up the samples. But the advantage of such a microcomputer is that a researcher is not dependent on a computer-centre, but can do work in a home environment at any time. With the programs developed it was possible to produce, for example, an overall TTR.

Table 3
Overall TTR

Variety	Types	Tokens	TTR
1. Daily newspapers	12,215	48,242	0.2532
2. Weekly newspapers	11,527	48,242	0.2389
3. Novels	10,090	48,242	0.2092
4. Women's weeklies	11,272	48,242	0.2337
5. Popular scientific prose	10,907	48,242	0.2261
6. Bureaucratic language	8,151	48,242	0.1690

What is remarkable here is the low TTR for bureaucratic language. Especially when one considers opinions in the literature on the positive relation between TTR and degree of education or readability of texts. But the data presented here are not reliable as the number of topics in bureaucratic language may be far smaller than in the other samples.

We can neutralize the effect of the number of topics when we look at function words, for example prepositions. But then we have to solve another problem, namely the difference in size or text length. The number of prepositions in these samples are different, as the following table shows under tokens:

Table 4
TTR of prepositions

Variety	Types	Tokens	TTR
1. Daily newspapers	63	6,132	0.0103
2. Weekly newspapers	61	5,635	0.0108
3. Novels	49	4,089	0.0120
4. Women's weeklies	52	5,066	0.0103
5. Popular scientific prose	56	5,752	0.0097
6. Bureaucratic language	66	7,819	0.0084

The number of prepositions is the lowest in novels and the highest in bureaucratic language. The high frequency of prepositions in bureaucratic language can be related to the results of the hypotheses discussed earlier concerning the nominal style of this kind of language. Prepositions are necessary if nouns are used in adjectival or adverbal appositions. In novels the opposite is the case: fewer nouns and prepositions. Uit den Boogaart (1975) gives 18.5 per cent of nouns against, for example, 23 per cent in popular scientific prose.

The results can be produced in graphs as in Figure 1; where the types are presented in diminishing frequency.

The y-axis gives the frequencies of the prepositions which are plotted on the x-axis, each width in the histogram standing for a preposition and each height for its frequency. The ten most frequent ones are shown here. Of note are the differences between the frequency of the most frequent preposition *van* (of). In bureaucratic language it occurs 2690 times, against, for example, 872 in novels. On the basis of this research on prepositions one could argue that a rough study of preposition frequency is not sufficient to characterize language varieties or LSPs. From the graphs one can see that the differences are mainly caused by the different frequencies of *van*.

We hope to describe these graphs by theoretical curves of the Zipf-Mandelbrot type. In this way each graph can be characterized by two numbers, one of which describes how fast the graph decreases to zero. This can be considered a measure of vocabulary richness. (See Gill and Renkema, forthcoming). With this example I hope to have given an outline of very detailed TTR research. Data on other word categories will be available in the near future.

To round off this survey on lexical parameters and computational linguistics, I should like to emphasize that bureaucratic language is only chosen as an example, and that the restriction to lexical parameters is arbitrary. Sentence level phenomena can be studied in the same way (see Renkema, 1981). But also in the study of other LSPs and other phenomena one has to start with the functions (no matter how difficult they are to describe) the LSP under consideration in fact has. From this functional or pragmatic starting point one has to choose other LSPs which can serve as a kind of related LSP, in order to lay one's finger on the characteristic phenomena. And furthermore it is very useful to study general parameters, for they can give a valuable insight into possibilities of variation, and so provide a contribution to the description of LSP in general.

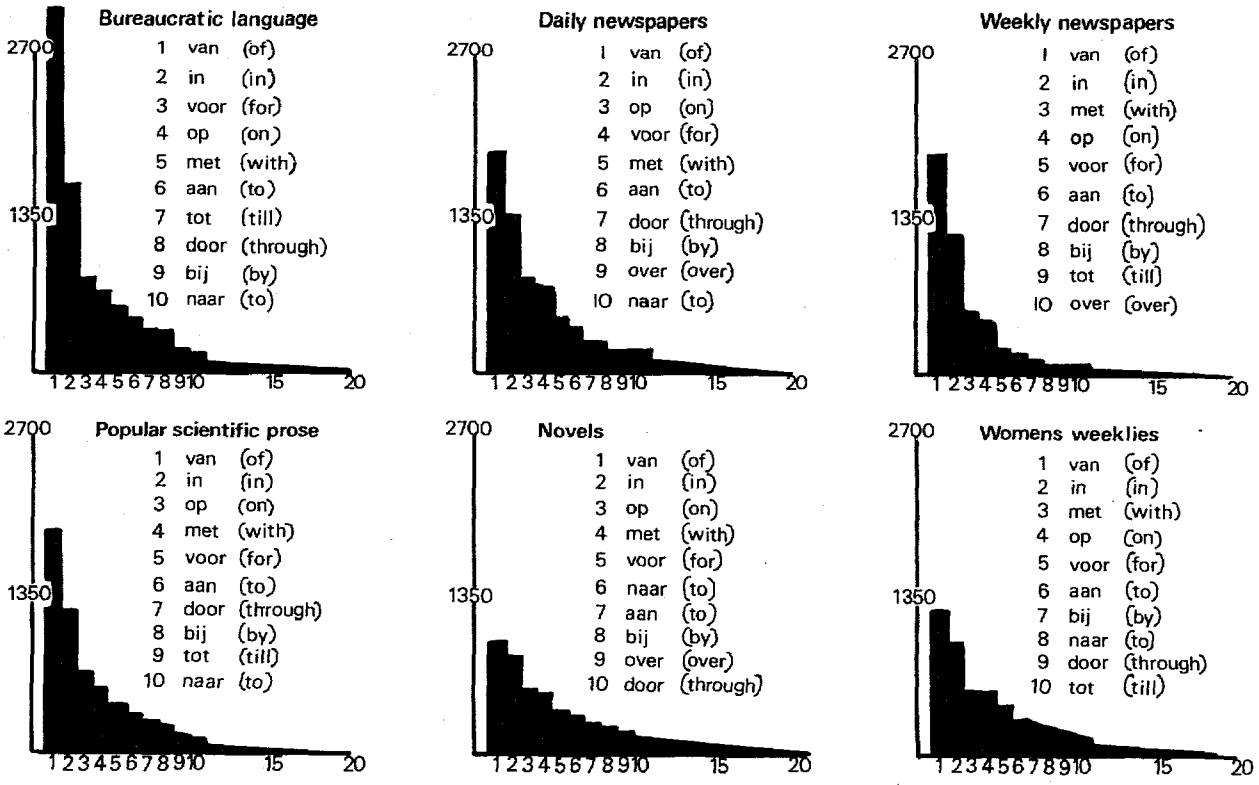

Figure 1. Preposition frequency in various types of prose.

Note

Thanks are due to J. Portier for her computational support and R.D. Gill for advice on statistics.

References

ANDROLINA, C. (1980) 'Syntactic maturity and vocabulary richness of learning-disabled children at four age levels'. *Journal of Learning Disabilities, 13*, pp. 373-7.

BRANDT CORSTIUS, H. (1978) *Computer-taalkunde*, Muiderberg: Coutinho.

CARROLL, J. (1964) *Language and Thought*. Englewood Cliffs, N.J.: Prentice-Hall.

GILL, R.D. and RENKEMA, J. (forthcoming) 'TTR and statistics'.

HERDAN, G. (1960) *Type-Token Mathematics. A Textbook of Mathematical Linguistics* 's-Gravenhage: Mouton.

JONG, E. de (ed.) (1979) *Spreektaal. Woordfrequenties in gesproken Nederlands*. Utrecht: Bohn, Scheltema & Holkema.

MANSCHRECK, T.C. MAHER, B.A. and ADER, D.N. (1981) 'Formal thought disorder, the type-token ratio, and disturbed voluntary motor movement in schizophrenia'. *British Journal of Psychiatry, 139*, pp. 7-15.

RENKEMA, J. (1977) 'Taal gebruik en spreiding van kennis'. *Onze Taal, 46*, pp. 26-30.

RENKEMA, J. (1981) *De taal van 'Den Haag': Een kwantitatief-stilistisch onderzoek naar aanleiding van oordelen over taalgebruik*. 's-Gravenhage: Staatsuitgeverij.

RENKEMA, J. (forthcoming) 'Type-token ratio, suggestions for an improved application'.

UIT DEN BOOGAART, P.C. (ed.) (1975) *Woordfrequenties in geschreven en gesproken Nederlands*. Utrecht: Bohn, Scheltema & Holkema.