Positional Occurrences in Texts: Weighted Consensus Strings

Peter Zörnig
Kamil Stachowski
Ioan-Iovitz Popescu
Tayebeh Mosavi Miangah
Ruina Chen
Gabriel Altmann

2016
RAM-Verlag

Studies in quantitative linguistics

Editors

Fengxiang Fan	(fanfengxiang@yahoo.com)
Emmerich Kelih	(emmerich.kelih@univie.ac.at)
Reinhard Köhler	(koehler@uni-trier.de)
Ján Mačutek	(jmacutek@yahoo.com)
Eric S. Wheeler	(wheeler@ericwheeler.ca)

1. U. Strauss, F. Fan, G. Altmann, *Problems in quantitative linguistics 1*. 2008, VIII + 134 pp.
2. V. Altmann, G. Altmann, *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen.* 2008, IV+193 pp.
3. I.-I. Popescu, J. Mačutek, G. Altmann, *Aspects of word frequencies*. 2009, IV +198 pp.
4. R. Köhler, G. Altmann, *Problems in quantitative linguistics 2.* 2009, VII + 142 pp.
5. R. Köhler (ed.), *Issues in Quantitative Linguistics.* 2009, VI + 205 pp.
6. A. Tuzzi, I.-I. Popescu, G.Altmann, *Quantitative aspects of Italian texts*. 2010, IV+161 pp.
7. F. Fan, Y. Deng, *Quantitative linguistic computing with Perl.* 2010, VIII + 205 pp.
8. I.-I. Popescu et al., *Vectors and codes of text*. 2010, III + 162 pp.
9. F. Fan, *Data processing and management for quantitative linguistics with Foxpro***.** 2010, V + 233 pp.
10. I.-I. Popescu, R. Čech, G. Altmann, *The lambda-structure of texts*. 2011, II + 181 pp
11. E. Kelih et al. (eds.), *Issues in Quantitative Linguistics Vol. 2*. 2011, IV + 188 pp.
12. R. Čech, G. Altmann, *Problems in quantitative linguistics 3*. 2011, VI + 168 pp.
13. R. Köhler, G. Altmann (eds.), *Issues in Quantitative Linguistics Vol 3.* 2013, IV + 403 pp.
14. R. Köhler, G. Altmann, *Problems in Quantitative Linguistics Vol. 4.* 2014, VIII+148 pp.
15. Best, K.-H., Kelih, E. (eds.), *Entlehnungen und Fremdwörter: Quantitative Aspekte.* 2014. VI + 163 pp.
16. I.-I. Popescu, K.-H. Best, G. Altmann, G. *Unified Modeling of Length in Language.* 2014, VIII + 123 pp.
17. G. Altmann, R. Čech, J. Mačutek, L. Uhlířová (eds.), *Empirical Approaches to Text and Language Analysis dedicated to Luděk Hřebíček on the occasion of his 80th birthday*. 2014. VI + 231 pp.

18. M. Kubát, V. Matlach, R. Čech, *QUITA Quantitative Index Text Analyzer*. 2014, VII + 106 pp.
19. K.-H. Best, *Studien zur Geschichte der Quantitativen Linguistik.* 2015. III + 158 pp.
20. P. Zörnig, K. Stachowski, I.-I. Popescu, T. Mosavi Miangah, P. Mohanty, E. Kelih, R. Chen, G. Altmann, *Descriptiveness, Activity and Nominality in Formalized Text Sequences.* 2015. IV+120 pp.
21. G. Altmann, *Problems in Quantitative Linguistics Vol. 5*. 2015. III+146 pp.
22. P. Zörnig, K. Stachowski, I.-I. Popescu, T. Mosavi Miangah, R. Chen, G. Altmann, *Positional occurrences in texts: Weighted Consensus Strings.* 2016. II+178 pp.

ISBN: 978-3-942303-37-8

RAM-Verlag
Stüttinghauser Ringstr. 44
D-58515 Lüdenscheid
Germany
RAM-Verlag@t-online.de
http://ram-verlag.de

Contents

1. Introduction

Written texts contain punctuation which allows us to mechanically determine units larger than the clause. Mostly such units represent some kind of grammatically determined sentences; other ones represent verses in a poem, written in one line. But poems may be written in such a way that sentences exceed the boundary of the verse. In that case one can analyze the poem in two ways. Spoken texts, e.g. telephone conversation, do not have any punctuation; one must determine "the sentence" either authoritatively or considering the intonation, or the change of the speaker in a stage play or some other signals.

If one analyzes the text, taking into account only special entities which occur in the predetermined frame-entities, one can perform a "consensus" analysis to be specified below either for the text as a whole and in turn, one can compare texts; or, if the sentences or verses are too short, one can determine rather parts of the text, e.g. Frumkina's sections, containing 100, 200,… words, or strophes, chapters, 10 sentences, etc. but the purpose of the analysis must be in some relation to this type of segmentation. Preliminarily, there is no prescription or a fixed way of constructing wholes/frames which should be sequentially analyzed.

Nevertheless, the text can always be transcribed either symbolically as a sequence of abbreviations of entities classified in some known way, e.g. parts of speech – abbreviated as *Art*, *Pn*, *Aj*, *Av*, *N*, *V*, *Pp*, *I*, *C*, etc. – or as degrees of properties of the individual entities, e.g. length, to obtain a sequence. Now, after having a sequence, there is a full pallet of statistical methods that can help us to state its properties. There are distances, transition frequencies, positional aspects, runs, etc. (see e.g. Zörnig et al. 2015).

In the present work we study some other aspects of a text which can be considered as a set of sequences written in separate lines, i.e. a text is an array

$$(1.1) \qquad \begin{pmatrix} s_1^{\,1} & s_2^{\,1} & \cdots \\ \vdots & \vdots & \vdots \\ s_1^{\,n} & s_2^{\,n} & \cdots \end{pmatrix} \quad ,$$

where the sequences (lines) $s^i = (s^i_1, s^i_2, \ldots)$ may have different lengths. We study the distribution of the elements in the *columns* of (1.1), in particular the most frequent element of a column is of great interest. In a certain sense we study a text "vertically", which is a novel approach to quantitative linguistics. We may compare and evaluate the columns and test whether there are some positional regularities. In some languages these are given already by syntactic rules. In poetry they may be prescribed by the rhythm or by positional assonances, in scientific texts one expects a certain ductus, and in stage plays there is a sequence of

speech acts, etc. In order to capture the positional occurrences we extend the concept of "consensus string", a term that has been recently transferred from computational biology to linguistics (Zörnig, Altmann 2016). A consensus string is a sequence $t = (t_1,\ldots,t_n)$ which is – in a sense to be concretized – as close as possible to the strings given in (1.1). One possibility to define $t = (t_1,\ldots,t_n)$ – which we adopt in linguistic applications – is setting t_j equal to one of the most frequent element of the j-th column of (1.1).

Definition 1: Let $\Sigma = \{s^1,\ldots, s^n\}$ be a set of sequences as in (1.1). Let $F(j)$ be the largest frequency of an element in column j and let $N(j)$ denote the number of elements in column j. The latter is equal to the number of sequences having length at least j. Then the *weighted consensus string* (WCS) of the set of sequences Σ is defined as the sequence $\left(\frac{F(j)}{N(j)}\right)_{j=1,2,\ldots}$

Irrespective of the type of the sequences s^i (which may be e.g. symbolic or numerical sequences), the WCS is always a uniquely determined numeric sequence.

Example 1: Consider the following 4 sentences:

s^1 = (1, 3, 2, 2, 4, 1),
s^2 = (1, 4, 3, 3, 2, 2, 2),
s^3 = (4, 1, 2, 3, 3, 3, 4, 2, 1),
s^4 = (2, 1, 2, 2, 1).

Since the strings are not equally long, one can bring them to the same lengths by adding zeroes, yielding

s^1 = (1, 3, 2, 2, 4, 1, 0, 0, 0),
s^2 = (1, 4, 3, 3, 2, 2, 2, 0, 0),
s^3 = (4, 1, 2, 3, 3, 3, 4, 2, 1),
s^4 = (2, 1, 2, 2, 1, 0, 0, 0, 0).

Thereby the strings have been made comparable, i.e. the Hamming distance or another distance (Zörnig, Altmann 2016, section 2) is now defined between any two of these strings.

One *possible* consensus string *CS* of the above example has the form

CS = (1, 1, 2, 3, 4, 1, 0, 0, 0).

This string is in general not uniquely determined. For a string matrix (1.1), any string having at position j a most frequent element of column j, is a consensus

string. For example, in column 4 of the above example there exist two most frequent elements, namely 2 and 3. Each of them could be the fourth element of CS. Clearly, different consensus strings have different distances to the given (observed) sequences.

Such a string $CS = (t_1,\ldots,t_m)$ minimizes the average distance to the given strings (Zörnig, Altmann 2016). The (uniquely defined) *weighted consensus string* is

WCS = (2/4, 2/4, 3/4, 2/4, 1/4, 1/3, 1/2, 1/1, 1/1).

In the following we will confine ourselves to weighted consensus strings.

Consider now some symbolic sequences:

Example 2: Given the five sequences

s^1 = (a, a, b, b, b, a, c, d, a, b, b)
s^2 = (a, b, d, a, a, c, d)
s^3 = (b, a, c, c, d, a, c, b, a, b)
s^4 = (d, c, b, a, b, a, c, d)
s^5 = (a, c, b, b, b, a, c, d, c, b, b)

over the alphabet {a, b, c, d}. For example, the first column contains 5 elements, where the most frequent one *a* occurs three times. Thus F(1) = 3, N(1) = 5. Column 4 contains 5 elements, the most frequent ones are *a* and *b*, occurring two times each. Thus F(4) = 2 (largest frequency) and N(4) = 5. The column 9 contains three elements and the most frequent is *a*, occurring 2 times. Thus, F(9) = 2, N(9) = 3. The WCS is therefore

(3/5, 2/5, 3/5, 2/5, 3/5, 4/5, 4/5, 3/4, 2/3, 3/3, 2/2)

For the purposes of the present book we do not need all complete strings of (1.1). It is sufficient to have a table of the following form.

Definition 2: Given the string matrix (1.1) where the elements of the strings are chosen from the alphabet $A = \{a_1,\ldots,a_k\}$. Then the table

	Columns			
	1	2	…	m
a_1 ⋮ a_k	$f_{1,1}$ ⋮ $f_{k,1}$	$f_{1,2}$ ⋮ $f_{k,2}$	… ⋮ …	$f_{1,m}$ ⋮ $f_{k,m}$
sum	N(1)	N(2)	…	N(m)

where $f_{i,j}$ denotes the frequency of the element a_i in the j-th column of (1.1), is called the *frequency table of the string matrix* (1.1). With the notations in Definition 1 it holds that $F(j)$ is the maximum value of column j and $N(j)$ is the sum of values in column j.

In the following chapters we express the information about a set of strings in form of its frequency table. For example, the frequency tables of the string matrices in Examples 1 and 2 are given by

Frequency table of Example 1

	Columns								
	1	2	3	4	5	6	7	8	9
1	2	2	0	0	1	1	0	0	1
2	1	0	3	2	1	1	1	1	0
3	0	1	1	2	1	1	0	0	0
4	1	1	0	0	1	0	1	0	0
sum	4	4	4	4	4	3	2	1	1

Frequency table of Example 2

	Columns										
	1	2	3	4	5	6	7	8	9	10	11
a	3	2	0	2	1	4	0	0	2	0	0
b	1	1	3	2	3	0	0	1	0	3	2
c	0	2	1	1	0	1	4	0	1	0	0
d	1	0	1	0	1	0	1	3	0	0	0
sum	5	5	5	5	5	5	5	4	3	3	2

Another way to characterize the columns is to consider

(a) The rank-frequency distributions of the individual columns and the parameters of the theoretical distribution.

(b) A function of the corresponding moments that shows a concentration to a certain structure. One can use also Ord's criterion, etc. These indicators compare different moments of the distribution but the testing of differences becomes more complex.

Though the above cases are merely examples, we may conjecture three hypotheses:

(1) Each column in the string matrix has its specific frequency distribution. The confirmation of this conjecture implies that something like a vertical structure of texts exists. Studying millions of sentences, we could find a distribution of sentence types viewed from the grammatical perspective. Any grammatical analysis is merely the stating of facts, not the finding of background

laws. These can be established only deductively but one cannot do it before one performs a lot of indictive work in many languages.

(2) The form of the consensus string depends on the stylistic homogeneity of the text.

(3) If the sentences do not have the same length, then the consensus string frequently increases, beginning at a position corresponding approximately to the median of the text lengths, i.e. a position for which 50% of the texts end before reaching that position.

The properties of the WCS can be measured by various indicators, e.g. using the Hurst exponent, the Minkowski sausage, the V indicator defined below, etc.

The hypotheses may have various boundary conditions (exceptions of the rule) according to the level of entities, kind of measurement, text type, spoken or written text, age of the author, etc. The only way to define them more exactly must be preceded by empirical investigations because up to now the investigation of this behavior of texts is not sufficiently known (cf. Hřebíček 2000; Zörnig, Altmann 2016).

The first hypothesis may have two forms: if the elements of the string are numbers or symbols, one may obtain the rank-frequency distribution. It is conjectured that both types of distributions can be derived from the unified theory (cf. Wimmer, Altmann 2005).

The second hypothesis merely says that there is variation in the frame units/strings. They mostly begin with an entity which is most frequent in the given language but afterwards they begin to vary. This need not be the case in poetry if it follows a special meter. Our aim is to find the kind of positional dependence of the function in the consensus string. In general, we may conjecture that some positions in the sentence are preferred by some type of entities while other ones are neglected. The WCS may be different according to the type of text.

The third hypothesis is quite evident: From a given point the number of zeroes increases, hence their proportion increases. For the sake of simplicity, we omit the zero positions.

In quantitative linguistics one strives for setting up a hypothesis and testing it. The hypothesis should be set up in such a way that it is statistically testable. Non-testable hypotheses are dogmas that cannot be used in science.

Here we test whether the weighted consensus string follows a certain law and expresses a text characteristic, a text type, a language, a development of a person etc. Our aim is to find the laws which may hold in different forms depending on the character of the considered entities, express them mathematically and use them for comparisons.

2. Units and Frames

The "vertical text analysis" introduced above has a large range of possible applications that increase with each further step into the depth of language.The elements of the sequences may in principle be linguistic units of any kind. In the present chapter we provide an overview of the most current ones. Table 2.1 contains a preliminary list of possibilities. Here, the framing entity can be either a sentence, a verse, a strophe, a person in a stage play, an act in a stage play, a chapter, Frumkina's sequence consisting of 50, 100, 200,… words, complete text, a text part containing 10 sentences, etc. The units of these frames can be given by numbers or symbols. The results will differ automatically by choosing a framing unit.

Table 2.1
Units and symbolization

Units/Properties	**Symbolization**
Word length	Discrete numbers
Morphological complexity of the word	Continuous numbers
Word frequency in text	Discrete numbers
Canonical syllable types	Symbols
Parts of speech	Symbols
Number of topical grammatical categories of the word (gender, number, case, time, mode,…)	Discrete numbers
Degree of word polysemy	Discrete numbers (from a dictionary)
Number of synonyms of a word	Discrete numbers (from a dictionary)
Number of verb valencies	Discrete numbers
Sentence length	Discrete numbers (in # of clauses)
Sentence complexity	Discrete numbers
Phrase types	Symbols
Length of topic	Discrete number
Length of comment	Discrete number
Psycholinguistic properties of the word	Symbol (e.g. dogmatism, imagery)
Canonical syllable types (in verse)	Symbol
Motifs of syllable types	Symbols
Rhythmic feet	0,1 or symbol (iamb, dactylus,…)
Rhythmic motifs	Symbols
Word length motifs	Discrete numbers

Word complexity motifs	Continuous numbers
Word frequency motifs	Discrete numbers
POS motifs	Symbols
Motifs of grammatical categories	Symbols
Motifs of polysemy	Discrete numbers
Motifs of synonymy	Discrete numbers
Speech acts	Symbols
Motifs of speech acts	Symbols
……..	……

The presented possibilities are only the best known ones. The list can be continued, specified, one can show the hierarchies, etc. Nevertheless, it may be a source of extensive investigations. It should be noted that definitions of linguistic entities are conventions. A different way of definition may lead to different results because the sets of entities are different. Different linguistic schools might tend to different opinions with respect to the above perspectives.

Our aim is to study some of the above aspects and test the hypotheses with preliminary data which could be improved or changed in the course of further investigations.

Of course, one can analyze any texts but the longer the texts are, the more meaningful are the results. Only sufficiently large text samples enable significant statements. Our aim is to find text-specific tendencies and, if possible, to propose a law that is valid for all texts in all languages.

We shall describe some of the aspects mentioned above. Some data can be found in the Appendix; here we show only the descriptions, tables with results, computations, figures, etc.

2.1. Word length

The word length is measured in terms of syllable numbers, i.e. based on phonetics and not on script which is not alphabetic in all languages and sometimes strongly redundant (e.g. Frech or English). In this way one can approach any living language – however, the dead languages may have problems because their phonetics are not always known. Nevertheless, one can measure word length also in terms of morpheme numbers but in strongly analytic languages one would obtain always the length 1.

It is more appropriate to evaluate word length in prose than in poetry where there may be prescriptions concerning word length. But poetry written in prose or simply poetry without meter can obey by the prosaic hypotheses. The results concerning “true” poetry may display a very special behavior.

Let us consider the word length in sentences in a Slovak prosaic text “Moja Dolná zem” by E. Bachletová. The individual strings can be found in the Appendix. Counting only numbers greater than zero we obtain the results pre-

sented in Table 2.1.1 (see Def. 2 in the introduction). Column *j* contains the frequencies of words of length *i*. For example, the second column contains 38 words of length 1, 31 words of length 2, etc. Sentences with more than 15 words were rare, thus we consider only the first 15 columns of the given text.

Table 2.1.1
Word length in sentence positions in the Slovak text by E. Bachletová

	Columns														
WL	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	30	38	22	15	20	15	14	16	11	4	6	7	4	3	6
2	30	31	37	34	27	21	24	21	14	13	11	10	10	8	4
3	21	12	16	19	17	21	9	9	13	15	13	9	3	4	3
4	8	6	7	6	7	3	6	7	6	4	5	1	2	3	2
5	3	1	1	1	-	2	2	-	1	2	2	1	1	1	-
6	-	-	-	-	1	-	-	-	1	1	-	-	-	-	-
Sum	92	88	83	75	72	62	55	53	46	39	37	28	20	19	15
WCS	.33	.43	.45	.45	.38	.34	.44	.4	.3	.38	.35	.36	.5	.42	.4

We suppose that there is a strict control and the frequency (probability) of length *x* is proportional to that of *x*−1. More precisely, we assume a simple proportionality between the neighboring classes which can be expressed by the proportionality function $g(x) = a/(b+x-1)$. In the numerator, there is merely a language constant *a* which is influenced by the effort of the hearer, *b*, and the topical length *x*−1. Thus we obtain the difference equation

$$(2.1) \quad P_x = \frac{a}{b+x-1} P_{x-1},$$

whose solution yields the Hyperpoisson distribution

$$(2.2) \quad P_x = \frac{a^x}{{}_1F_1(1;b;a)b^{(x)}}$$

where ${}_1F_1(1;b;a) = 1 + a/b + a^2/[b(b+1)] + \ldots$ denotes the hypergeometric function and $b^{(x)} = b(b+1)\ldots(b+x-1)$ the rising factorial. Fitting this distribution to the column values in Table 2.1.1 we obtain the results presented in Table 2.1.2. Since we consider merely non-zero lengths, the distribution is displaced one step to the right, i.e. it begins with $x = 1$. The software automatically pools classes in which $NP < 1.0$, where *N* denotes the sample size.

Table 2.1.2
Fitting the Hyperpoisson distribution to the columns in Table 2.1.1
(Slovak: Bachletová)

Length x	Columns, computed values				
	1	2	3	4	5
1	29.24	37.50	21.77	15.14	19.52
2	31.99	30.29	35.99	34.31	28.11
3	19.31	14.12	18.33	18.47	16.39
4	8.05	4.63	5.52	5.64	5.99
5	3.40	1.46	1.40	1.43	1.60
6	-	-	-	-	0.40
a	1.3472	1.1029	0.7360	0.7061	0.9800
b	1.2315	1.365	0.4452	0.3115	0.6805
DF	2	2	2	2	2
X^2	0.34	0.89	0.84	0.17	0.75
P	0.84	0.64	0.66	0.92	0.69

Length x	Columns, computed values				
	6	7	8	9	10
1	13.91	14.37	15.18	10.53	3.84
2	24.51	21.12	19.92	15.80	14.17
3	15.51	12.84	11.78	11.48	12.26
4	5.98	4.92	6.11	5.50	6.01
5	2.09	1.75	-	1.97	2.05
6	-	-	-	0.73	0.68
a	0.9871	1.0370	1.0765	1.4077	1.1295
b	0.5603	0.7055	0.8202	0.9383	0.3057
DF	2	2	1	2	2
X^2	4.02	1.82	0.89	0.65	1.41
P	0.13	0.40	0.35	0.72	0.49

Length x	Columns, computed values				
	11	12	13	14	15
1	4.73	6.7541	3.57	3.04	5.18
2	13.64	11.4450	8.92	7.17	4.88
3	11.08	6.6947	5.26	5.39	2.96
4	5.24	2.3665	1.76	2.41	1.98
5	2.31	0.7397	0.50	1.00	-
a	1.1307	0.8933	0.7721	1.1028	1.7024
b	0.3917	0.5272	0.3088	0.4678	1.8050
DF	2	1	1	1	1
X^2	1.24	1.38	1.40	0.56	0.29
P	0.54	0.24	0.24	0.46	0.59

Further testing for the Slovak text (columns ≥ 16) is irrelevant here because either the values become too small or the number of degrees of freedom becomes insufficient. The results show that in our case the word lengths in the first 15 sentence positions follow the Hyperpoisson distribution.

The data suggest that the parameters *a* and *b* are correlated. If one orders the data according to increasing *a*, one obtains the results presented in Table 2.1.3. Fitting the power function $y = kx^c$, one obtains $b = 0.5856a^{1.8249}$ with the determination coefficient $R^2 = 0.88$.

Table 2.1.3
Link between parameters *a* and *b*

a	**b**	**b computed**
0.7061	0.3115	0.3103
0.7360	0.4452	0.3347
0.7721	0.3088	0.3653
0.8920	1.1081	0.4754
0.8933	0.5272	0.4766
0.9800	0.6805	0.5644
0.9871	0.5603	0.5719
1.0370	0.7055	0.6257
1.0765	0.8202	0.6699
1.1028	0.4678	0.7001
1.1295	0.3057	0.7313
1.1307	0.3917	0.7327
1.3472	1.2315	1.0088

1.4077	0.9383	1.0930
2.5634	3.2997	3.2632
k = 0.5856; c = 1.8249; R^2 = 0.878		

The result proves to be satisfactory. The word length in individual positions in Slovak sentences follows a conjectured law. The Hyperpoisson distribution expresses the positional dependence of word length, and the zeta-function shows the very regular course of the Hyperpoisson in the subsequent columns. The dependence of *b* on *a* is shown in Figure 2.1.1.

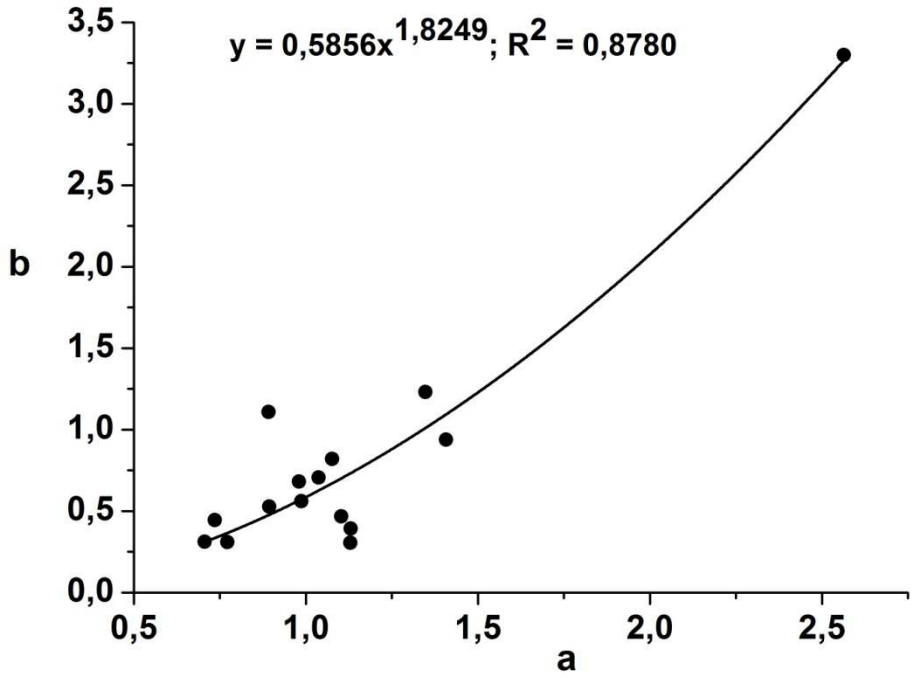


Figure 2.1.1. The link between the parameters of the Hyperpoisson distribution in the Slovak text: $b = f(a)$

The interpretations of local extremes and the shape of the WCS may depend on the regarded entities. Let us consider some other languages. The results for 5 Persian texts are presented in Table 2.1.4 to 2.1.8.

Table 2.1.4
Length in Persian: Text 1

	Column											
Length	1	2	3	4	5	6	7	8	9	10	11	12
1	8	12	8	10	11	7	9	6	8	7	12	6
2	7	5	8	9	8	6	3	7	6	8	7	8
3	7	9	10	8	9	12	13	8	10	10	6	11
4	4	1	1	3	2	2	1	4	4	2	4	4
5	2	1	2			2	3	3	1	1	0	1
6	1	1	1			0	1	0	1	1	0	
7	1	1				1		0		1	0	
8								2			1	
Sum	30	30	30	30	30	30	30	30	30	30	30	30
WCS	.27	.40	.33	.33	.37	.40	.43	.27	.33	.33	.40	.37

	Column									
Length	13	14	15	16	17	18	19	20	21	22
1	11	11	10	9	10	6	8	7	9	3
2	5	7	7	5	4	5	2	4	5	4
3	8	3	6	5	4	5	4	7	5	6
4	5	5	4	2	4	4	3	3	0	2
5	1	1	1	2	2	1	3		1	4
6		0		1		1	1		1	0
7		0					1			1
8		1								
Sum	30	28	28	24	24	22	22	21	21	20
WCS	.37	.39	.36	.38	.42	.27	.36	.33	.43	.30

Table 2.1.5
Length in Persian: Text 2

	Column											
Length	1	2	3	4	5	6	7	8	9	10	11	12
1	9	4	10	8	14	4	7	9	9	3	12	9
2	5	9	5	5	5	4	7	4	10	11	2	9
3	6	6	4	8	7	7	10	12	6	7	9	8
4	5	6	8	4	1	5	3	2	4	5	5	1
5	3	2	1	4	1	4	3	2	1	3	2	0
6	0	1	0	0	0	0		0		1		1

7	1	1	1	0	1	2		0				1
8	1	1	1	0	1	4		0				1
9				0				0				
10				1				1				
Sum	30	30	30	30	30	30	30	30	30	30	30	30
WCS	.30	.30	.33	.27	.47	.23	.33	.40	.33	.37	.40	.30

	Column								
Length	13	14	15	16	17	18	19	20	21
1	12	11	9	13	10	11	8	8	6
2	9	7	4	7	6	2	8	2	3
3	5	2	8	6	3	2	3	6	4
4	2	5	5	1	3	5	0	3	2
5	1	2	2	1	3	1	2	2	4
6	1	1	1		1	1	0	0	0
7		2				2	1	1	0
8							1		1
Sum	30	30	29	28	26	24	23	22	20
WCS	0.40	.37	.31	.46	.38	.46	.35	.36	.30

Table 2.1.6
Length in Persian: Text 3

	Column											
Length	1	2	3	4	5	6	7	8	9	10	11	12
1	5	6	3	4	10	10	7	12	10	9	9	9
2	4	7	12	8	4	9	5	5	3	2	5	8
3	12	10	7	13	7	6	6	7	9	7	3	4
4	5	7	4	2	5	3	9	5	5	6	7	6
5	1		3	1	3	1	1	1	1	3	4	0
6	1		0	1	1	1	1		1	1	1	2
7	0		0	1			1			1		
8	1		1									
9	1											
Sum	30	30	30	30	30	30	30	30	29	29	29	29
WCS	.40	.33	.40	.43	.33	.33	.30	.40	.34	.31	.31	.31

	Column								
Length	13	14	15	16	17	18	19	20	21
1	11	10	10	9	7	6	6	7	6
2	5	2	6	10	9	6	5	5	4
3	11	6	5	4	8	6	3	7	8
4	1	9	5	4	2	6	6	0	0
5	1	2	1	0		1	3	2	1
6			1	1		1			2
7			1						
Sum	29	29	29	28	26	26	23	21	21
WCS	.38	.34	.34	.36	.35	.23	.26	.33	.38

Table 2.1.7
Length in Persian: Text 4

	Column											
Length	1	2	3	4	5	6	7	8	9	10	11	12
1	8	6	2	10	5	14	12	9	9	5	14	9
2	5	9	10	6	8	6	3	9	7	8	3	5
3	7	5	12	5	8	7	9	6	10	10	9	4
4	5	8	2	3	6	1	2	4	2	6	1	8
5	3	2	4	4	1	2	2	2	0	0	3	2
6	1			0	1		1		1	1		0
7	0			1	0		0		0			0
8	1			0	1		0		0			0
9				0			1		0			1
10				1					1			
Sum	30	30	30	30	30	30	30	30	30	30	30	29
WCS	.27	.30	.40	.33	.27	.47	.40	.30	.33	.33	.47	.31

	Column						
Length	13	14	15	16	17	18	19
1	9	8	8	11	9	6	7
2	8	7	5	5	7	11	6
3	6	6	8	5	4	4	2
4	2	4	3	2	3	0	1
5	2		1	0		1	2
6	0			1		1	2
	1						1
Sum	28	25	25	25	23	23	21
WCS	.32	.32	.32	.44	.39	.48	.33

Table 2.1.8
Length in Persian: Text 5

	Column								
Length	1	2	3	4	5	6	7	8	9
1	5	8	4	11	11	5	20	5	15
2	4	12	5	4	3	9	2	8	5
3	10	1	12	11	4	8	2	5	4
4	7	5	7	4	9	4	3	5	3
5	4	1	2		1	2	2	3	0
6		1			1	0		1	0
7		0			0	0		0	1
8		0			1	0		0	
9		1				1		1	
10		1							
SUM	30	30	30	30	30	29	29	28	28
WCS	.33	.40	.40	.37	.37	.31	.69	.29	.54

	Column				
Length	10	11	12	13	14
1	6	8	11	8	8
2	10	6	5	7	3
3	9	8	7	3	6
4	1	3	2	4	3
5	1	2	0	1	3
6	0		2	0	1
7	0			1	
8	1				
Sum	28	27	27	24	24
WCS	.36	.30	.41	.33	.33

The course of the WCS of length in the five Persian texts is displayed in Figure 2.1.2.

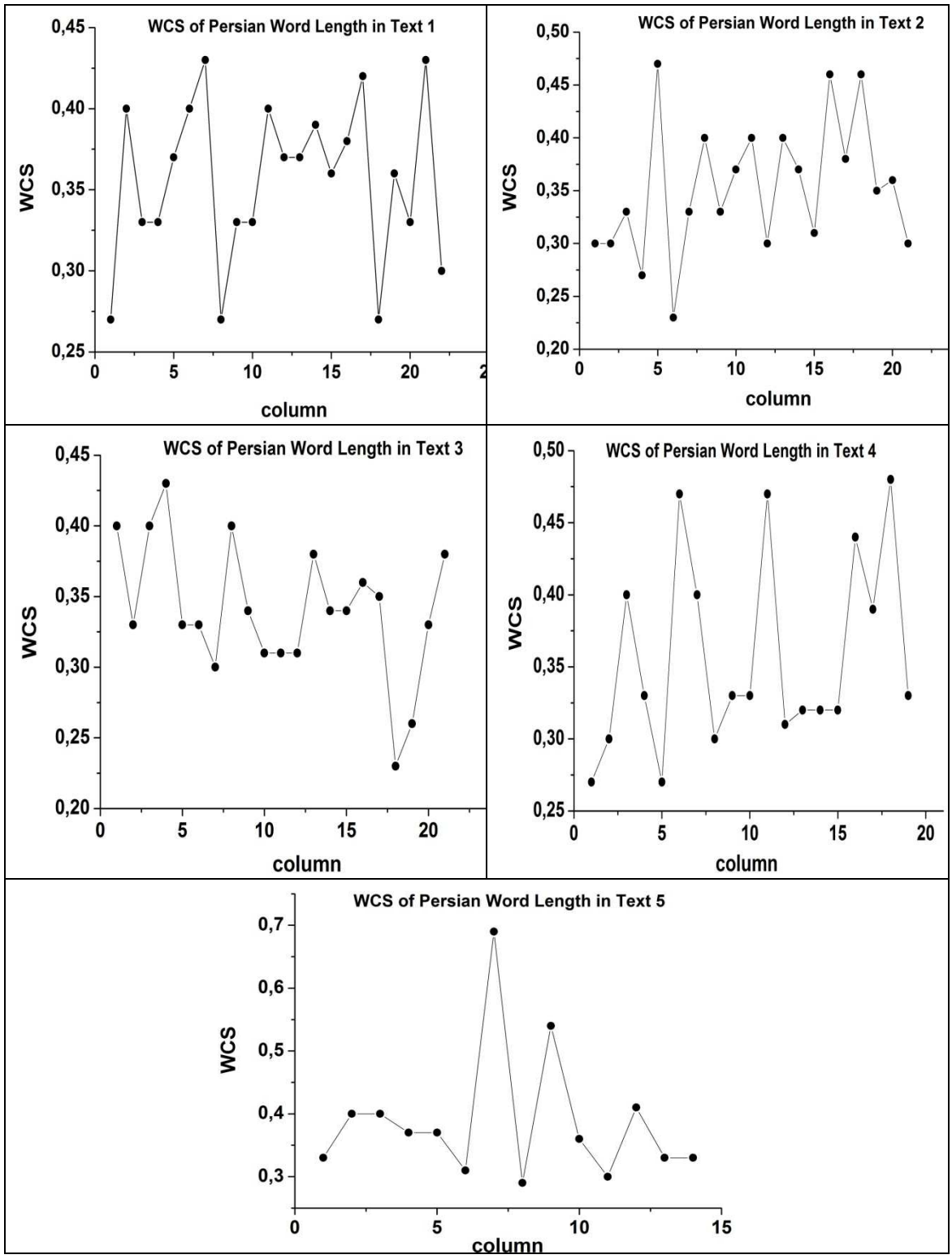


Figure 2.1.2. The course of WCS of length in Persian texts

The fitting of the Hyperpoisson distribution to the length-frequency of words in columns is a preliminary attempt. One may obtain good/better results using e.g. the distributions Hirata-Poisson, Cohen-Poisson, Ferreri-

Poisson, or Singh-Poisson, or their zero-truncated (positive) variants (Wimmer, Altmann 1999). But here we shall maintain the Hyperpoisson distribution and introduce another one which is also adequate for other data. Thus in Text 1, column 5 and 7 will be checked separately. The results for the five Persian texts are presented in Tables 2.1.9 to 2.1.13.

Table 2.1.9
Fitting the Hyperpoisson distribution to the frequency of word lengths in Persian: Text 1

	Column							
L	1	2	3	4	6	8	9	10
1	7.74	10.53	7.35	9.96	7.31	6.47	6.75	7.55
2	7.77	7.85	9.19	10.23	8.77	7.54	9.45	8.28
3	6.19	5.22	7.05	6.11	6.91	6.57	7.38	6.58
4	4.08	3.13	3.90	3.70	4.05	4.57	4.00	4.09
5	2.34	1.71	1.69		1.89	2.64	1.66	2.09
6	1.13	0.86	0.84		0.73	1.31	0.76	0.91
7	0.78	0.68			0.34	0.57		0.50
8						0.33		
a	3.8439	6.1717	1.9840	1.4285	2.2928	3.4421	1.7688	2.8703
b	3.8268	8.2814	1.5867	1.3906	1.9107	2.9504	1.2634	2.6150
DF	3	3	2	1	3	3	2	3
X^2	0.24	5.87	3.69	0.86	5.69	0.52	2.49	3.72
P	0.97	0.12	0.16	0.35	0.13	0.91	0.29	0.29

	Column							
L	11	12	13	14	15	16	17	18
1	12.42	5.54	9.18	10.83	8.86	8.65	8.07	6.83
2	8.20	10.58	8.91	7.19	8.73	6.18	6.45	5.70
3	4.80	8.13	6.21	4.47	5.80	4.04	4.40	4.10
4	2.52	3.91	3.38	2.62	2.90	2.43	2.61	2.59
5	1.20	1.85	2.33	1.44	1.71	1.36	2.47	1.46
6	0.53			0.76		1.33		1.32
7	0.21			0.37				
8	0.12			0.32				
a	5.1611	1.2852	2.4736	9.8636	2.0380	7.7751	4.6253	5.1765
b	7.8209	0.6726	2.5491	14.8627	2.0683	10.8942	5.7839	6.1961
DF	2	2	2	3	2	3	2	3
X^2	1.90	2.07	4.13	2.94	1.20	0.93	2.26	1.37
P	0.39	0.35	0.13	0.40	0.55	0.82	0.32	0.71

	Column			
Length	19	20	21	22
1	3.95	6.78	8.71	8.71
2	5.54	6.28	5.36	5.36
3	5.20	4.22	3.16	3.16
4	3.67	3.71	1.79	1.79
5	2.07		0.97	0.97
6	0.98		1.00	1.00
	0.60			
a	2.8417	2.4544	13.9907	13.9907
b	2.0266	2.6502	22.7138	22.7138
DF	3	1	2	2
X^2	7.34	2.80	2.23	2.23
P	0.06	0.09	0.33	0.33

Table 2.1.10
Fitting the Hyperpoisson distribution to the frequency
of word lengths in Persian: Text 2

	Column						
Length	1	2	3	4	6	7	9
1	8.59	4.91	7.60	5.59	3.58	6.72	8.98
2	6.89	7.19	6.97	7.98	4.53	8.91	9.98
3	5.16	7.06	5.61	7.27	4.93	7.22	6.52
4	3.62	5.22	4.03	4.86	4.71	4.21	3.02
5	2.38	3.09	2.62	2.57	4.00	2.94	1.50
6	1.48	1.53	1.56	1.12	3.06		
7	0.87	0.65	0.85	0.42	2.14		
8	1.00	0.35	0.77	0.14	3.04		
9				0.04			
10				0.01			
a	11.0875	2.9817	6.7347	2.5187	7.7635	2.0807	1.5890
b	13.8139	2.0362	7.3583	1.7649	6.1354	1.5682	1.4301
DF	4	3	4	3	5	2	2
X^2	2.15	1.37	8.31	3.48	4.37	1.84	0.53
P	0.71	0.71	0.08	0.32	0.50	0.40	0.77

	Column						
Length	10	12	13	15	16	17	18
1	3.39	8.94	11.98	7.78	12.69	7.61	6.69
2	9.19	8.84	8.60	7.81	8.46	5.79	5.49
3	8.92	6.28	5.09	6.05	4.28	4.17	4.17
4	5.26	3.48	2.56	3.82	1.75	2.86	2.95
5	2.24	1.56	1.12	2.03	0.83	1.89	1.95
6	1.00	0.61	0.65	1.51		1.16	1.22
7		0.20				0.69	1.52
8		0.08				0.84	
a	1.5101	2.5165	3.3640	3.3892	2.1000	13.7847	10.3395
b	0.5571	2.5444	4.6845	3.3743	3.1500	18.1121	12.5985
DF	2	2	2	3	1	4	4
X^2	1.01	2.36	0.17	3.22	1.07	2.88	8.21
P	0.60	0.31	0.92	0.36	0.30	0.58	0.08

	Column		
Length	19	20	21
1	7.12	7.52	5.81
2	5.20	5.50	4.54
3	3.68	3.72	3.35
4	2.53	2.35	2.35
5	1.69	1.38	1.57
6	1.09	0.77	1.00
7	0.69	0.76	0.61
8	1.00		0.77
a	22.7839	9.0919	13.7500
b	31.1740	12.4276	17.6000
DF	4	3	3
X^2	4.67	4.29	1.60
P	0.32	0.23	0.66

Table 2.1.11
Fitting the Hyperpoisson distribution to the frequency
of word lengths in Persian: Text 3

		Column							
L	1	2	3	4	5	6	8	9	10
1	4.81	5.08	5.96	4.71	9.07	10.02	10.77	8.92	8.54
2	7.67	9.31	8.75	9.43	7.56	9.02	8.35	7.89	6.85
3	7.29	8.11	7.49	8.32	5.53	5.92	5.40	5.66	5.06
4	4.94	7.51	4.53	4.71	3.60	3.07	3.00	3.42	3.47

5 6 7 8	2.60 1.12 0.41 0.17		2.11 0.80 0.26 0.09	1.96 0.65 0.23	2.12 2.11	1.30 0.68	2.47	1.78 1.33	2.22 1.33 1.52
a	2.3548	1.6594	2.0539	1.5779	5.9751	2.4259	3.9350	3.8119	9.4614
b	1.4779	0.9049	1.3983	0.7890	7.1701	2.6955	5.0774	4.3129	11.800
DF	3	1	3	2	3	2	2	3	4
X^2	5.84	1.22	3.16	4.53	3.66	0.00	4.16	6.29	6.57
P	0.12	0.27	0.37	0.10	0.30	1.00	0.13	0.10	0.16

	Column							
Length	11	12	15	16	17	18	19	20
1	7.85	7.45	9.66	9.29	5.85	5.44	5.36	6.69
2	6.82	8.49	7.20	8.88	11.12	7.17	5.63	6.96
3	5.30	6.53	4.96	5.65	6.32	6.17	4.75	4.39
4	3.72	3.78	3.17	2.70	2.72	3.94	3.34	1.99
5	2.39	1.76	1.89	1.03		2.00	3.93	0.97
6	2.92	1.00	1.06	0.45		1.28		
7			1.06					
a	7.4223	2.3532	8.9238	1.9100	0.8107	2.4806	4.2652	1.6053
b	8.5532	2.0631	11.9625	1.9993	0.4267	1.8832	4.0601	1.5432
DF	3	2	4	2	1	3	2	1
X^2	6.87	2.84	1.70	1.41	1.27	1.89	3.13	2.43
P	0.08	0.24	0.79	0.49	0.26	0.60	0.21	0.12

Table 2.1.12
Fitting the Hyperpoisson distribution to the frequency
of word lengths in Persian: Text 4

	Column						
Length	1	2	3	4	5	6	7
1	7.66	5.67	1.94	9.60	5.03	13.48	10.92
2	7.27	8.50	9.71	6.93	8.05	7.95	7.51
3	5.85	7.51	9.83	4.83	7.58	4.37	4.87
4	4.08	4.71	5.53	3.26	5.06	2.25	2.99
5	2.52	3.60	2.99	2.12	2.61	1.94	1.74
6	1.39			1.34	1.10		0.97
7	0.70			0.82	0.39		0.51
8	0.53			0.49	0.16		0.25
9				0.28			0.23
10				0.33			
a	5.2857	2.1507	1.2683	19.9908	2.2877	8.0533	11.4941
b	5.5714	1.4338	0.2537	27.6805	1.4298	13.6467	16.7150

DF	4	2	2	4	3	2	3
X^2	1.40	3.91	3.08	2.46	1.29	2.77	6.67
P	0.84	0.14	0.21	0.65	0.74	0.25	0.08

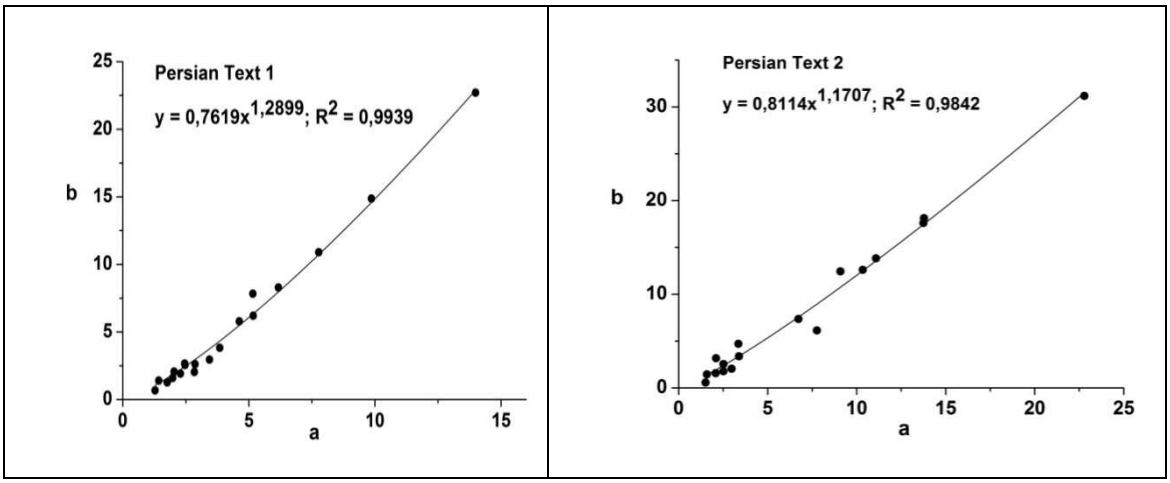
	Column						
Length	8	9	10	12	13	14	15
1	9.10	9.86	4.50	8.94	9.09	7.74	6.15
2	9.00	7.93	7.20	8.84	7.88	7.81	8.53
3	6.27	5.50	7.88	6.28	5.42	5.25	6.04
4	3.37	3.35	6.56	3.48	3.09	4.21	2.87
5	2.25	1.82	4.40	1.56	1.51		1.41
6		0.89	4.46	0.61	0.64		
7		0.40		0.20	0.36		
8		0.16		0.08			
9		0.06					
10		0.03					
a	2.3516	5.0072	3.4633	2.5165	3.3395	2.0149	1.4462
b	2.3778	6.2246	2.1645	2.5444	3.8529	1.9978	1.0421
DF	2	3	3	2	2	1	2
X^2	0.16	6.37	5.69	2.36	0.55	0.21	2.78
P	0.92	0.10	0.13	0.31	0.76	0.65	0.25

	Column			
Length	16	17	18	19
1	10.09	8.97	5.77	7.37
2	6.91	6.97	10.58	5.05
3	3.91	4.07	5.00	3.34
4	1.88	2.99	1.35	2.13
5	0.79		0.27	1.32
6	0.43		0.04	0.79
7				1.00
a	3.2376	2.3392	0.6361	18.6772
b	4.7270	3.0075	0.3469	27.2509
DF	2	1	1	3
X^2	0.96	0.002	0.30	2.51
P	0.62	0.99	0.59	0.47

Table 2.1.13
Fitting the Hyperpoisson distribution to the frequency of word lengths in Persian: Text 5

	Column								
L	1	3	6	8	10	11	12	13	14
1	0.94	3.96	4.97	5.75	6.50	7.89	9.82	8.19	7.37
2	7.70	8.56	8.94	6.85	10.83	7.93	7.66	6.26	5.89
3	10.26	8.38	7.79	6.12	6.93	5.70	4.88	4.10	4.29
4	7.44	5.31	4.48	4.38	2.74	3.19	2.63	2.35	2.82
5	3.70	3.79	1.92	2.61	0.78	2.29	1.23	1.19	1.72
6	1.40		0.66	1.33	0.18		0.78	0.55	1.94
7	0.57		0.19	0.60	0.03			0.36	
8			0.05	0.24	0.01				
9			0.01	0.12					
a	1.5903	1.7936	1.6881	3.5787	1.0376	2.5188	3.4872	4.5532	7.6879
b	0.1934	0.8306	0.9378	3.0026	0.6225	2.5039	4.4707	5.9518	9.6154
DF	2	2	2	3	1	2	2	2	3
X^2	0.07	4.42	0.07	0.63	0.87	1.45	2.14	0.57	3.59
P	0.96	0.11	0.97	0.88	0.35	0.49	0.34	0.75	0.31

Again, the parameters *a* and *b* are correlated by the power relation as shown in Figure 2.1.3.



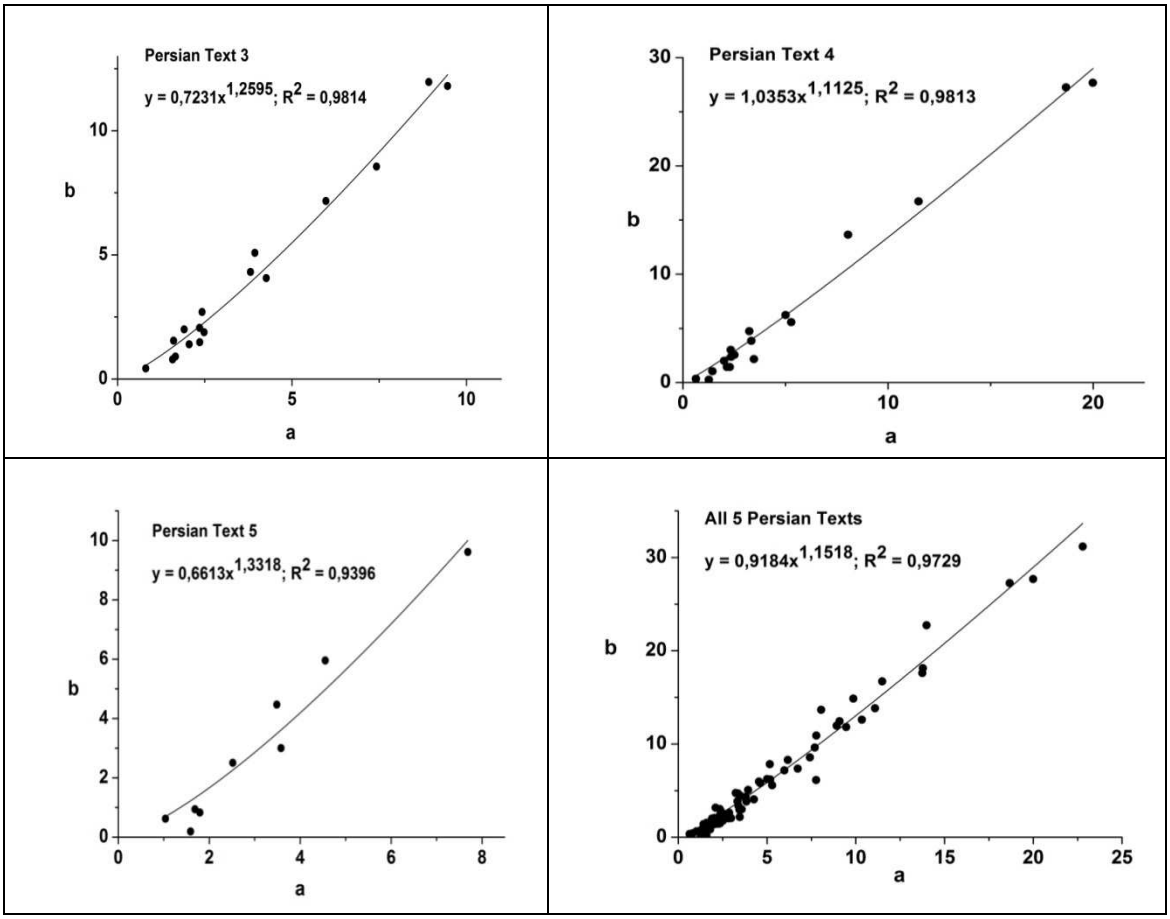


Figure 2.1.3. The relation between *a* and *b* in Persian length

The distributions (columns) that could not be satisfactorily fitted by the Hyperpoisson distribution are presented in Table 2.1.14 and 2.1.15. We omitted all those with a chi-square value less than 0.05 and those in which the parameters *a* and *b* assumed extreme values. The Hyperpoisson distribution has been chosen because of its easy interpretability.

On empirical reasons we fitted the exceptions by the Cohen-Poisson distribution, in which the first two values are modified. The distribution has been shifted one step to the right because the values begin with $x = 1$ and use the formula

$$(2.3)\quad P_x = \begin{cases} e^{-a}(1+\alpha a), & x = 0 \\ ae^{-a}(1-\alpha), & x=1 \\ \dfrac{a^x e^{-a}}{x!}, & x = 2,3,\ldots \end{cases}$$

Table 2.1.14
Persian texts: 1-displaced Cohen-Poisson

	Column						
Length	T1 C5	T1 C7	T2 C5	T2 C8	T2 C11	T3 C13	T3 C14
1	11.02	8.22	13.17	8.36	9.98	10.21	7.77
2	8.86	4.82	6.76	5.84	4.52	6.91	3.46
3	6.50	8.10	6.48	8.00	7.96	7.02	7.84
4	3.61	5.12	2.58	4.71	4.60	3.29	5.41
5		2.43	0.77	2.08	2.94	1.57	4.52
6		1.31	0.18	0.73			
7			0.04	0.22			
8			0.01	0.05			
9				0.01			
10				0.003			
a	1.1995	1.8967	1.1950	1.7651	1.7328	1.4048	2.0703
α	0.1828	0.4355	0.3771	0.3557	0.5087	0.3091	0.5427
DF	1	3	1	3	2	2	2
X^2	1.76	7.25	0.60	4.19	2.28	4.64	5.48
P	0.18	0.06	0.44	0.24	0.32	0.10	0.06

	Column				
Length	T3 C21	T4 C11	T5 C2	T5 C4	T5 C5
1	7.05	10.74	7.72	10.15	7.57
2	5.78	5.21	1.00	6.25	4.06
3	4.95	7.71	7.72	7.61	8.11
4	2.22	4.06	6.39	5.99	5.59
5	0.74	2.28	3.97		2.89
6	0.26		1.97		1.20
7			0.82		0.41
8			0.29		0.16
9			0.09		
10			0.03		
a	1.3441	1.5791	2.4836	1.5335	2.0688
α	0.2142	0.4661	0.8392	0.3702	0.4818
DF	1	2	3	1	3
X^2	2.61	4.68	5.60	3.05	7.26
P	0.11	0.10	0.13	0.08	0.06

In three cases, neither the Hyperpoisson nor the Cohen-Poisson distributions yielded satisfactory results. One could capture them using another modified Poisson distribution, namely the Singh-Poisson defined as

$$(2.4)\quad P_x = \begin{cases} 1-\alpha+\alpha e^{-a}, & x=0 \\ \dfrac{\alpha a^x e^{-a}}{x!}, & x=1,2,3,... \end{cases}$$

and shifted one step to the right. The results are presented in Table 2.1.15.

Table 2.1.15
Fitting the Singh-Poisson distribution to some Persian data

		Column		
Length	T 2 C 14	T 3 C 7	T 5 C 7	T 5 C 9
1	10.30	6.30	18.92	14.88
2	4.77	4.33	2.73	5.12
3	5.68	6.62	3.03	4.24
4	4.52	6.74	2.24	2.35
5	2.69	5.15	2.08	0.97
6	1.28	3.15		0.32
7	0.75	2.70		0.12
a	2.3846	3.0556	2.2172	1.6577
α	0.7232	0.8604	0.3902	0.5788
DF	3	4	2	2
X^2	4.16	7.99	0.87	0.32
P	0.24	0.09	0.65	0.85

The length in **Polish** data is displayed in Tables 2.1.16 to 2.1.20.

Table 2.1.16
Word length in Polish texts: Text 1

	Column										
Length	1	2	3	4	5	6	7	8	9	10	11
1	5	6	3	4	10	10	7	12	10	9	9
2	4	7	12	8	4	9	8	5	3	2	6
3	12	10	7	13	7	6	6	7	9	7	3
4	5	7	4	2	5	3	9	5	5	6	7
5	1		3	1	3	1	1	1	1	3	3
6	0		0	1	1	1	1		0	1	1

7 8	1 1		0 1	1			1		1	1	
Sum	30	30	30	30	30	30	30	30	29	29	29
WCS	.4	.33	.4	.43	.33	.33	.3	.4	.34	.31	.31

		Column								
Length	12	13	14	15	16	17	18	19	20	21
1 2 3 4 5 6 7	9 8 4 6 0 2	11 5 11 1 1	10 2 6 9 2	10 6 5 5 1 1 1	9 10 4 4 0 1	7 9 8 2	6 6 6 6 1 1	6 5 3 6 3	7 5 7 0 2	6 4 8 0 1 2
Sum	29	29	29	29	28	26	26	23	21	21
WCS	.31	.38	.34	.34	.36	.35	.23	.26	.33	.38

Table 2.1.17
Word length in Polish texts: Text 2

	Column										
Length	1	2	3	4	5	6	7	8	9	10	11
1 2 3 4 5 6 7 8 9	15 16 12 7	14 18 8 6 4	17 12 7 9 3 0 2	10 12 11 13 3	11 18 10 6 3 1	11 8 17 8 4	9 14 10 11 3	15 8 14 6 3	12 15 9 5 2 0 1 0 1	9 14 10 7 1 1 1	17 8 13 4 1
Sum	50	50	50	49	49	48	47	46	45	43	43
WCS	.32	.36	.34	.24	.37	.35	.30	.33	.33	.33	.40

		Column			
Length	12	13	14	15	16
1	4	8	5	8	6
2	8	8	4	4	6
3	14	10	11	4	3
4	7	8	8	11	5

5	4	3	2	2	2
6	3	1	0		
7	1		1		
8			1		
Sum	41	38	32	29	22
WCS	.34	.26	.34	.38	.27

Table 2.1.18
Word length in Polish texts: Text 3

	Column											
Length	1	2	3	4	5	6	7	8	9	10	11	12
1	11	12	12	18	7	11	11	9	9	9	8	8
2	15	13	7	10	16	16	16	14	10	12	9	12
3	9	18	13	12	17	11	12	9	9	9	14	9
4	13	3	15	6	8	7	3	7	9	6	3	3
5	1	3	3	3	1	3	3	4	1	1	1	0
6	1	1						1	3			0
7												1
SUM	50	50	50	49	49	48	45	44	41	37	35	33
WCS	.30	.36	.30	.37	.35	.33	.36	.32	.24	.32	.40	.36

	Column				
Length	13	14	15	16	17
1	5	8	5	1	2
2	8	6	10	8	7
3	7	7	6	7	7
4	0	5	2	5	4
5	9	3	2	3	1
6	1				
7	1				
Sum	31	29	25	24	21
WCS	.29	.28	.40	.33	.33

Table 2.1.19
Word length in Polish texts: Text 4

	Column											
Length	1	2	3	4	5	6	7	8	9	10	11	12
1	20	19	19	16	22	14	13	10	15	13	14	19
2	11	13	8	19	11	21	19	12	14	12	11	11
3	11	11	15	12	11	8	13	14	9	5	11	12
4	7	5	6	3	6	2	1	8	6	11	6	2
5	0	2	1			3	2	3	1	0	0	1
6	1		1			1	1		1	4	3	
Sum	50	50	50	50	50	49	49	47	46	45	45	45
WCS	.40	.38	.38	.38	.44	.43	.39	.30	.33	.29	.31	.42

	Column								
Length	13	14	15	16	17	18	19	20	21
1	9	8	16	8	8	7	8	12	8
2	12	11	13	15	11	12	8	10	6
3	10	11	6	9	9	10	6	3	12
4	10	7	3	6	7	4	7	5	3
5	3	3	2	1	2	2	5	3	2
6	1	2	1			1	1	1	1
Sum	45	42	41	39	37	36	35	34	32
WCS	.27	.26	.39	.38	.30	.33	.23	.35	.38

Table 2.1.20
Word length in Polish texts: Text 5

	Column										
Length	1	2	3	4	5	6	7	8	9	10	11
1	10	15	15	8	12	12	12	7	14	9	12
2	21	16	20	16	17	15	18	19	15	14	10
3	13	11	9	19	9	11	9	13	4	8	8
4	6	4	6	5	5	7	6	2	7	7	4
5		2			3	1	1	2		1	2
6		1									
7		1									
Sum	50	50	50	48	46	46	46	43	40	39	36
WCS	.42	.32	.40	.40	.37	.33	.39	.44	.38	.36	.33

		Column			
Length	12	13	14	15	16
1	12	8	4	4	2
2	10	10	11	8	8
3	5	7	10	8	8
4	4	5	2	2	2
5	1			1	
Sum	32	30	27	23	20
WCS	.38	.33	.41	.35	.40

The Polish texts reveal the same trend as the Persian ones. However, the situation is here more diverse. For the sake of good fitting we shall choose always the most adequate model among the Hyperpoisson (HP), Cohen-Poisson (CP) and Singh-Poisson (SP) distributions. As already said above, the choice depends on the boundary conditions which are conjectured but not yet known. The parameter α belongs either to the Cohen-Poisson or to the Singh-Poisson distributions as shown in the second row of the tables. Usually, all of these distributions can be fitted satisfactorily but determining adequate boundary conditions will be possible only after many languages have been analyzed. The simple Poisson or positive Poisson distributions could be used in many cases but we shall not apply them here. The results of adjustment are presented in Tables 2.1.21 to 2.1.25.

Table 2.1.21
Fitting theoretical distribution to Polish word length: Text 1

	Column						
Length	1	2 CP	3	4	5	6	7
1	5.27	5.39	5.55	4.97	9.07	9.95	6.20
2	7.60	9.14	9.07	9.34	7.56	9.04	8.45
3	7.02	7.96	7.81	8.17	5.53	5.96	7.76
4	4.77	7.51	4.56	4.65	3.60	3.08	5.38
5	2.59		2.02	1.96	2.12	1.31	2.99
6	1.14		0.72	0.66	2.11	0.67	1.39
7	0.43		0.21	0.24			0.83
8	0.20		0.07				
a	2.5680	1.7288	1.8181	1.6345	5.9751	2.3914	2.8156
b	1.7789	-	1.1124	0.8701	7.1701	2.6310	2.0655
α	-	0.0075	-	-	-	-	-
DF	3	1	2	2	3	2	3
X^2	6.24	1.13	2.59	4.76	3.66	0.003	4.31
P	0.10	0.29	0.27	0.09	0.30	0.999	0.23

	Column						
Rank	8	9	10	11	12	13*	14
1	10.79	9.62	8.54	7.96	9.06	8.94	9.66
2	8.35	7.90	6.85	7.15	8.29	10.35	7.20
3	5.40	5.30	5.06	5.53	5.79	6.20	4.96
4	3.00	3.20	3.47	3.76	3.30	2.51	3.17
5	2.47	1.73	2.22	2.28	1.58	1.00	1.89
6		0.84	1.33	2.34	1.00		1.06
7		0.61	1.52				1.06
8							
a	3.9350	4.9425	9.4614	5.5883	3.0280	1.2433	8.9238
b	5.0774	6.1781	11.7996	6.2242	3.3192	1.0737	11.9625
α	-	-	-	-	-	-	-
DF	2	3	4	3	2	1	4
X^2	4.16	6.91	6.57	5.27	2.91	7.60	1.90
P	0.13	0.07	0.16	0.15	0.23	0.006 *	0.79

In column 13, there is a frequency distribution which cannot be fitted by any of the simple models. As can be seen, the observed values are bimodal, a phenomenon which can be observed also in several other columns of Text 1 but could be captured by the Hyperpoisson distribution.

Table 2.1.22a
Fitting theoretical distribution to Polish word length: Text 2

	Column				
Length	1	2	3	5	7
1	11.23	13.60	17.17	11.47	6.23
2	19.64	15.82	13.30	16.01	15.32
3	12.52	11.31	8.96	11.90	13.74
4	6.61	5.84	5.34	6.14	7.53
5		3.43	2.85	2.39	4.18
6			1.38	1.00	
7			1.00		
a	1.0024	1.8559	5.1500	1.6162	1.4105
b	0.5731	1.5953	6.6461	1.1573	0.5735
DF	1	2	3	2	2
X^2	1.98	1.38	3.14	0.71	4.30
P	0.16	0.50	0.37	0.70	0.12

	Column						
Rank	8	9	10	12	13	14	16
1	12.15	12.83	10.20	3.48	5.76	4.55	4.99
2	14.12	13.77	13.69	9.77	11.01	7.93	6.16
3	10.44	9.56	10.39	11.50	10.31	8.09	5.10
4	5.66	4.91	5.49	8.57	6.40	5.84	3.18
5	3.63	2.00	2.23	4.67	2.97	3.26	2.58
6		0.67	0.74	2.01	1.55	1.49	
7		0.26	0.26	1.00		0.57	
8						0.26	
a	2.0346	1.9672	1.7466	2.0294	1.8365	2.4679	2.5157
b	1.7513	1.8327	1.3017	0.7231	0.9609	1.4175	2.0381
DF	2	2	2	3	3	3	2
X^2	4.66	0.20	0.59	1.65	2.30	4.37	2.25
P	0.10	0.90	0.74	0.65	0.51	0.22	0.32

Table 2.1.22b
Fitting the Cohen-Poisson distribution to Polish word length: Text 2

	Column		
Rank	4	6	11
1	8.01	8.68	15.68
2	14.31	14.67	10.76
3	13.16	12.72	10.08
4	8.00	7.30	4.48
5	5.52	4.63	2.00
a	1.8247	1.7218	1.3336
α	0.0073	0.0070	0.2879
DF	2	2	2
X^2	5.50	5.25	2.22
P	0.06	0.07	0.33

Column 15 of text 2 very “irregular” and cannot be captured by any simple distribution

Table 2.1.23
Fitting the Hyperpoisson distribution to Polish word length: Text 3

	Column						
Length	1	2	4	5	6	7	8
1	9.61	11.74	16.16	0.67	9.67	10.34	8.47
2	14.84	16.00	14.32	22.44	16.49	16.04	13.11
3	12.74	12.15	9.62	17.15	12.58	11.25	11.22
4	7.57	6.40	5.20	6.63	6.19	5.09	6.64

5	3.44	2.58	3.71	2.11	3.08	2.28	3.00
6	1.79	1.14					1.55
a	1.9323	1.7147	2.7736	0.7822	1.3813	1.2795	1.9150
b	1.2515	1.2578	3.1299	0.0233	0.8100	0.8250	1.2374
DF	3	3	2	1	2	2	3
X^2	7.27	5.27	2.36	0.87	0.51	1.18	1.08
P	0.06	0.15	0.31	0.35	0.78	0.55	0.78

	Column					
Rank	9	10	11	12	14	16
1	7.65	7.03	5.69	9.56	6.22	0.30
2	10.68	13.79	14.02	11.61	8.29	7.62
3	9.79	9.95	9.82	7.34	6.96	8.56
4	6.68	4.40	4.00	3.14	4.26	4.92
5	3.63	1.84	1.46	1.01	3.27	2.60
6	2.57			0.26		
7				0.07		
a	2.6689	1.1418	0.9777	1.3206	2.2722	1.1745
b	1.9109	0.5823	0.3968	1.0881	1.7066	0.0457
DF	3	2	2	2	2	1
X^2	3.13	1.84	4.92	0.74	1.29	0.49
P	0.37	0.40	0.09	0.69	0.53	0.48

Columns 3 and 13 cannot be fitted by the Hyperpoisson distribution. Modifications require a further parameter.

Table 2.1.24
Fitting the Hyperpoisson distribution to Polish word length: Text 4

	Column						
Length	1	2	3	4	5	6	7
1	18.94	17.56	17.38	12.51	19.45	15.31	13.81
2	14.75	15.69	14.540	23.31	15.20	15.85	17.17
3	8.96	9.68	9.460	10.75	8.87	10.33	11.08
4	4.46	4.56	5.03	3.44	6.47	4.91	4.82
5	1.88	2.50	2.27			1.84	1.58
6	1.00		1.32			0.76	0.53
a	2.7657	1.9943	2.9285	0.6131	2.3048	1.7589	1.3402
b	3.5520	2.2316	3.5015	0.3290	2.9487	1.6990	1.0779
DF	2	2	3	1	1	2	2
X^2	4.15	0.90	7.31	1.97	2.04	4.79	3.98
P	0.13	0.64	0.06	0.16	0.15	0.09	0.14

	Column						
Rank	8	9	11	12	13	14	15
1	6.50	14.58	12.77	17.22	7.09	7.05	15.85
2	16.43	14.55	12.08	15.04	13.48	11.96	11.78
3	13.77	9.56	9.09	8.18	12.26	10.97	7.14
4	6.92	4.68	5.69	3.23	7.32	6.89	3.66
5	3.37	1.83	3.04	1.32	3.26	3.29	1.62
6		0.81	2.33		1.60	1.83	0.95
a	1.2543	1.9212	3.6885	1.4410	1.7418	1.9959	3.2879
b	0.4964	1.9250	3.8992	1.6500	0.9156	1.1761	4.4242
DF	2	2	3	2	3	3	2
X^2	3.29	0.59	3.86	3.60	2.32	0.25	0.50
P	0.19	0.74	0.28	0.17	0.51	0.97	0.78

	Column				
Rank	16	17	18	19	20
1	6.76	5.94	7.03	6.37	11.49
2	15.43	12.70	11.91	8.80	8.95
3	10.75	10.41	9.47	8.22	6.10
4	4.42	5.28	4.92	5.80	3.69
5	1.64	2.66	1.90	3.29	2.01
6			0.77	2.52	1.75
a	1.0031	1.3308	1.4981	2.8851	5.4283
b	0.4394	0.6231	0.8838	2.0875	6.9673
DF	2	2	2	3	3
X^2	1.34	1.85	0.24	3.14	2.99
P	0.51	0.40	0.89	0.37	0.39

Table 2.1.25
Fitting the Hyperpoisson distribution to Polish word length: Text 5

	Column						
Length	1	2	3	5	6	7	8
1	6.40	15.50	13.34	11.74	9.59	11.20	7.20
2	24.32	14.71	20.09	15.65	17.26	17.68	19.69
3	13.90	10.22	11.37	10.92	11.95	11.14	11.41
4	5.38	5.60	5.20	5.16	5.14	4.38	3.70
5		2.53		2.53	2.08	1.60	1.00
6		0.98					
7		0.46					
a	0.6729	2.5930	0.9090	1.4640	1.1295	1.0490	0.7356
b	0.1771	2.7329	0.6021	1.0980	0.6287	0.6649	0.2690
DF	1	3	1	2	2	2	1

X^2	2.61	0.98	0.83	0.55	2.21	1.29	0.35
P	0.11	0.81	0.36	0.76	0.33	0.52	0.55

	Column							
Rank	9	10	11	12	13	14	15	16
1	12.69	7.49	10.92	11.32	5.69	2.09	2.64	2.40
2	13.16	14.28	11.44	10.14	11.80	14.24	10.15	8.65
3	8.40	10.45	7.69	6.18	8.03	7.88	6.87	5.95
4	5.80	4.73	3.81	2.85	4.48	2.79	2.55	3.00
5		2.05	2.15	1.50			0.80	
a	1.6511	1.1871	1.8756	1.9058	1.0123	0.6022	0.8214	0.8504
b	1.5861	0.6222	1.7897	2.1274	0.4879	0.0884	0.2136	0.2362
DF	1	2	2	2	1	1	1	1
X^2	2.96	2.51	0.32	0.90	1.41	3.28	1.38	1.16
P	0.09	0.28	0.85	0.64	0.24	0.07	0.24	0.28

Column 4 does not follow the Hyperpoisson distribution but it can be satisfactorily modeled by the right truncated Poisson

$$(2.5)\quad P_x = \frac{a^x}{x!F(R)},\quad x = 0,1,...,R$$

where R is the truncation point and $F(R)$ is the sum of all frequencies. The distribution must be displaced one step to the right. We obtain the values

Length	1	2	3	4
f_x	8	16	19	5
NP_x	9.08	16.12	14.32	8.48

and the parameters are a = 1.7763, R = 4, DF = 1, X^2 = 3.09, P = 0.08. Evidently, there is some boundary condition influencing the usual course of frequencies.

The WCS of word length in Polish texts is displayed in Figure 2.1.4.

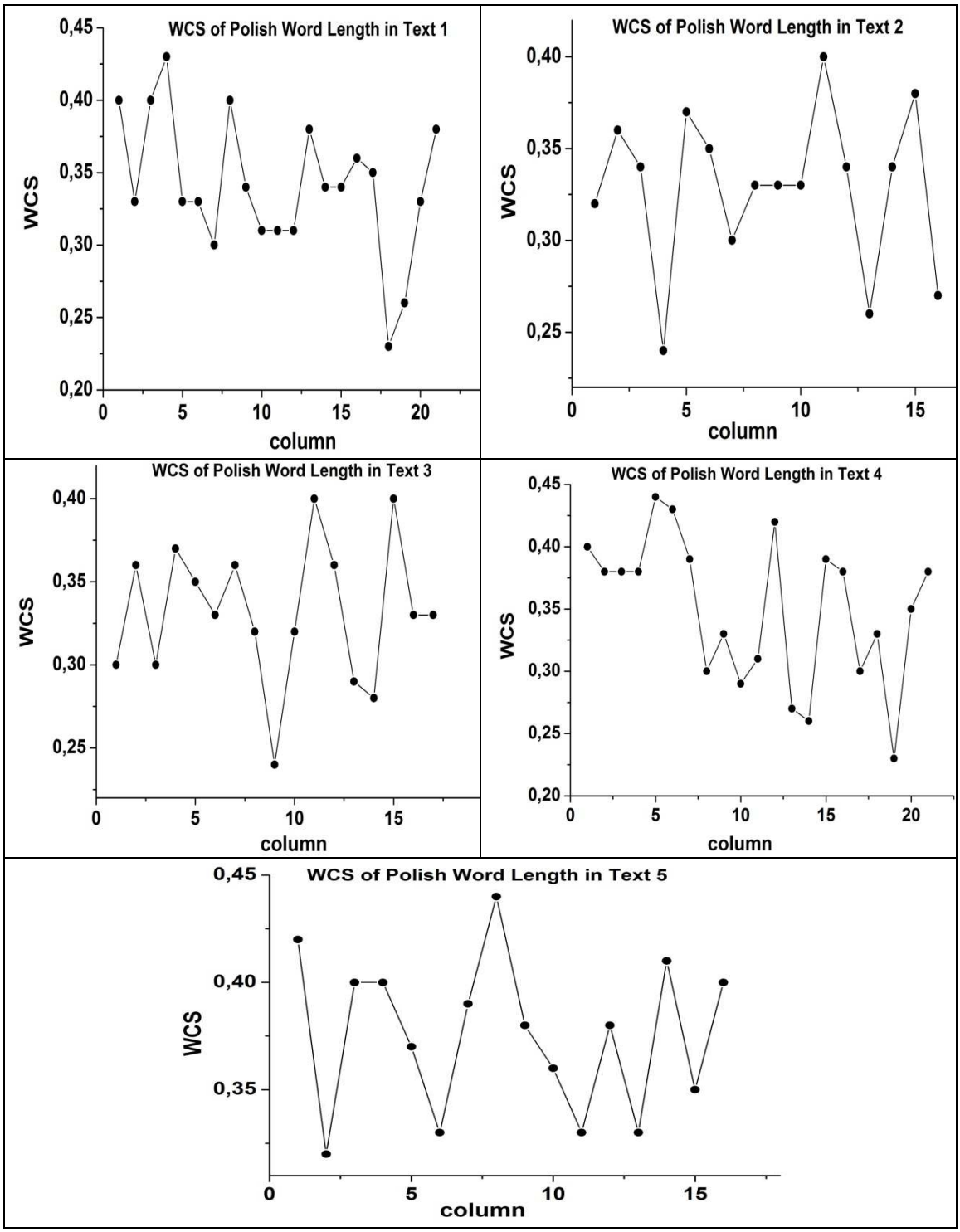


Figure 2.1.4. Polish word length WCS for the 5 texts

The distributions of **Turkish** word lengths are presented in Tables 2.1.26 to 2.1.30

Table 2.1.26
Turkish word length: Text 1

	Column						
ength	1	2	3	4	5	6	7
1	12	9	10	6	11	5	8
2	26	24	16	14	11	13	9
3	7	8	19	14	8	11	12
4	3	4	4	9	10	11	6
5	2	3		3	0	0	4
6		1			1	1	
7		1					
SUM	50	50	49	46	45	41	39
WCS	.52	.48	.39	.30	.24	.32	.31

	Column				
Length	8	9	10	11	12
1	8	3	4	2	3
2	15	12	9	15	6
3	8	7	8	2	9
4	4	6	4	4	3
5	1	3	0	0	
6			1	1	
SUM	36	31	26	24	21
WCS	.42	.39	.35	.62	.43

Table 2.1.27
Turkish word length: Text 2

	Column						
Length	1	2	3	4	5	6	7
1	9	5	5	3	4	2	4
2	17	15	13	13	12	11	7
3	10	8	17	8	12	10	6
4	8	10	3	5	6	5	5
5	4	7	3	4	1	2	4
6	1	2	5	4	1	1	1
7	1	2	1	2			2
8		1					
SUM	50	50	47	39	36	31	29
WCS	.34	.30	.36	.33	.33	.35	.24

	Column			
Length	8	9	10	11
1	4	4	3	2
2	6	6	8	3
3	5	8	6	5
4	4	0	1	6
5	5	2	1	3
6	1	2	2	0
7	1	1		0
1				1
SUM	26	23	21	20
WCS	.23	.35	.38	.30

Table 2.1.28
Turkish word length, Text 3

	Column						
Length	1	2	3	4	5	6	7
1	12	8	7	9	4	7	7
2	12	10	18	13	19	12	12
3	10	17	10	11	11	8	12
4	13	10	9	13	8	14	10
5	2	2	3	1	2	4	3
6	1	3	2	1	2	0	0
7					1	1	1
SUM	50	50	49	48	47	46	45
WCS	.26	.34	.37	.27	.40	.30	.27

	Column							
Length	8	9	10	11	12	13	14	15
1	6	7	6	7	4	8	9	4
2	15	12	9	10	8	9	9	7
3	12	14	15	8	11	11	5	11
4	4	3	5	8	7	4	5	4
5	3	2	3	1	2	1	2	2
6		1		1	1			
SUM	40	39	38	35	33	33	30	28
WCS	.38	.36	.39	.29	.33	.33	.30	.39

Table 2.1.29
Turkish word length, Text 4

	Column						
Length	1	2	3	4	5	6	7
1	11	10	11	9	11	11	10
2	10	15	14	12	14	13	10
3	13	11	11	18	7	11	15
4	6	9	10	8	10	6	10
5	8	3	3	3	5	7	2
6	2	1	1		0		
7		0			1		
8		1					
SUM	50	50	50	50	48	48	47
WCS	.26	0.30	.28	.36	.29	.27	.32

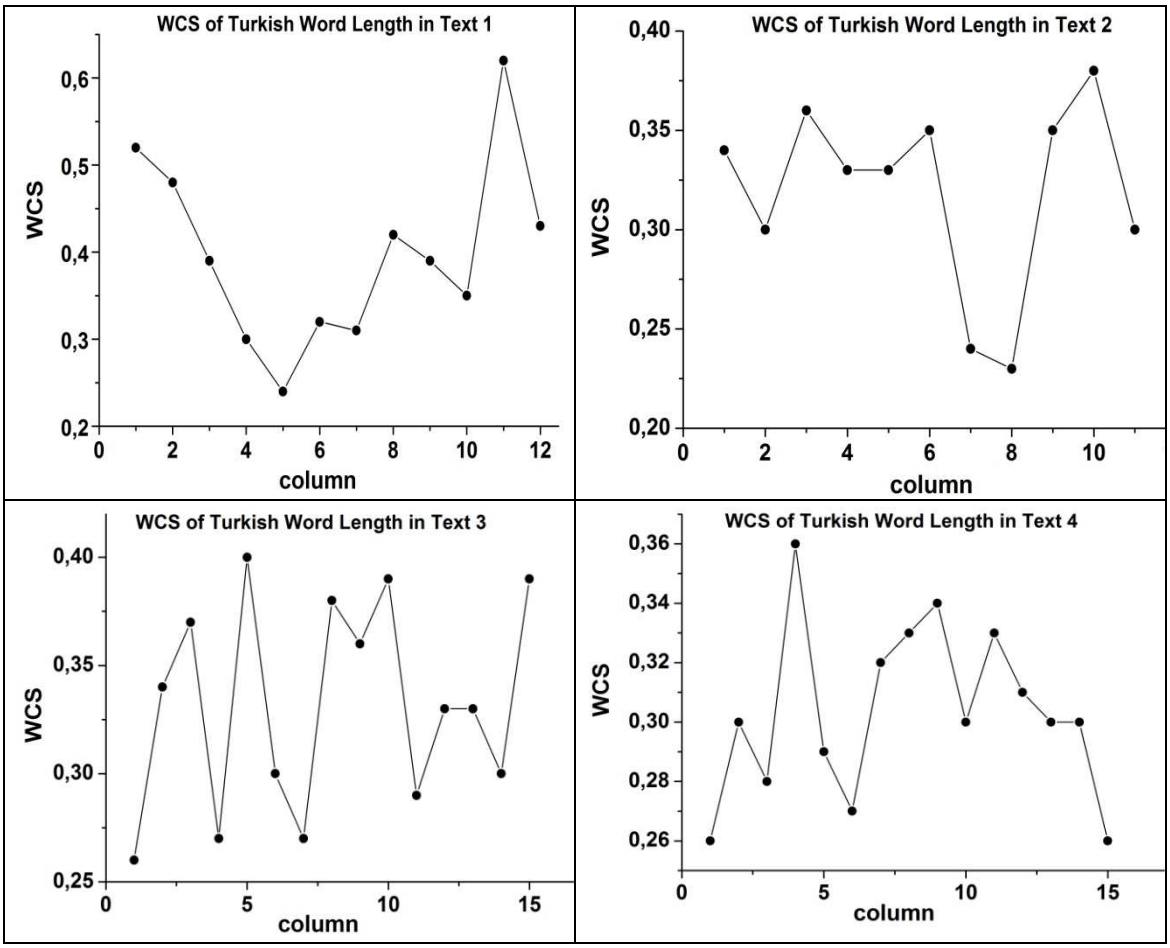
	Column							
Length	8	9	10	11	12	13	14	15
1	10	8	11	7	5	5	4	6
2	13	11	11	13	5	7	5	5
3	15	15	12	8	11	9	8	6
4	4	7	5	5	9	4	3	4
5	3	3	1	5	5	4	4	2
6	1			1		1	1	
SUM	46	44	40	39	35	30	27	23
WCS	.33	.34	.30	.33	.31	0.30	.30	.26

Table 2.1.30
Turkish word length: Text 5

	Column						
Length	1	2	3	4	5	6	7
1	12	13	9	6	6	8	6
2	12	15	10	13	8	14	17
3	12	11	16	17	12	9	7
4	10	8	11	9	17	11	8
5	3	2	2	2	3	1	3
6	1	1	2	3	1	1	3
7					1	1	
SUM	50	50	50	50	48	45	44
WCS	.24	.30	.32	.34	.35	.31	.39

	Column							
Length	8	9	10	11	12	13	14	15
1	11	9	8	7	6	4	4	3
2	8	5	10	7	3	9	7	4
3	6	16	8	8	13	10	6	4
4	7	6	6	7	5	5	2	8
5	4	1	2	2	3	1	2	3
6	6	3	2	1	1		2	0
7	1		1	2			1	1
1				1				
SUM	43	40	37	35	31	29	24	23
WCS	.26	.40	.27	.23	.42	.34	.29	.35

The WSC of Turkish word-lengths is presented in Figure 2.1.5.



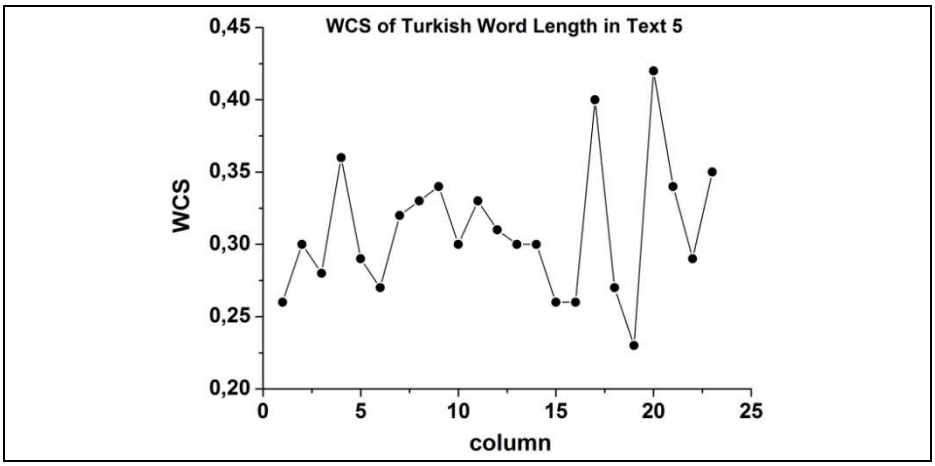


Figure 2.1.5. WCS of Turkish word lengths in 5 texts

The fitting of the Hyperpoisson distribution to Turkish word lengths is presented in Tables 2.1.31 to 2.1.35

Table 2.1.31
Fitting the Hyperpoisson distribution to Turkish word lengths: Text 1

	Column					
Length	1	2	3	4	5	6
1	11.08	12.76	11.43	5.37	9.82	5.22
2	24.02	16.28	18.28	15.28	11.72	13.56
3	11.29	11.77	12.16	13.87	9.39	12.13
4	2.97	5.93	7.13	7.49	5.67	6.55
5	0.63	2.30		4.00	2.75	2.53
6		0.72			1.65	1.00
7		0.25				
a	0.5999	1.6678	1.1376	1.3330	2.4454	1.3633
b	0.2769	1.3069	0.7110	0.4684	2.0505	0.5243
DF	1	2	1	2	3	2
X^2	2.41	7.53	5.69	0.74	6.70	4.98
P	0.12	0.02	0.02	0.69	0.08	0.08

	Column					
Length	7	8	9	10	11	12
1	8.23	7.80	2.85	3.98	1.57	2.78
2	11.04	14.62	10.42	9.56	11.75	7.88
3	9.37	9.08	9.76	7.46	7.50	6.24
4	5.81	3.38	5.25	3.47	2.50	4.10
5	4.55	1.12	2.72	1.15	0.56	
6				0.37	0.11	
a	2.3121	0.9281	1.2600	1.1558	0.6978	1.1006

b	1.7248	0.4950	0.3450	0.4816	0.0930	0.3884
DF	2	2	2	2	1	1
X^2	1.20	0.27	1.17	0.33	6.10	1.98
P	0.55	0.87	0.56	0.85	0.01	0.16

Three columns in Text 1, namely 2, 3 and 11 yield very low *P*-values. We could test other distributions, but we want to postpone this until boundary conditions are known.

Table 2.1.32
Fitting the Hyperpoisson distribution to Turkish word lengths: Text 2

	Column					
Length	1	2	3	4	5	6
1	9.31	6.25	5.89	4.61	3.62	2.08
2	14.69	10.80	11.10	8.42	13.08	10.60
3	12.82	11.87	11.94	9.42	11.28	10.20
4	7.73	9.56	8.98	7.60	5.52	5.42
5	3.56	6.08	5.20	4.79	1.89	1.99
6	1.33	3.19	2.44	2.48	0.62	0.71
7	0.56	1.43	1.43	1.70		
8		0.83				
a	1.9513	3.0171	2.5058	2.8856	1.1322	1.1864
b	1.2366	1.7451	1.3302	1.5786	0.3133	0.2328
DF	3	4	4	4	2	2
X^2	1.06	3.99	10.32	5.28	0.32	0.09
P	0.79	0.41	0.04	0.26	0.85	0.96

	Column				
Length	7	8	9	10	11
1	3.90	3.72	4.17	3.96	1.62
2	6.73	5.58	5.39	6.20	4.25
3	7.12	5.85	5.11	5.38	5.28
4	5.43	4.72	3.82	3.22	4.29
5	3.24	3.09	2.36	1.47	2.59
6	1.59	1.70	1.24	0.77	1.24
7	1.00	1.34	0.92		0.50
8					0.24
a	2.7328	3.4892	3.5388	1.9365	2.3486
b	1.5824	2.3261	2.7347	1.2350	0.8919
DF	3	4	3	2	3
X^2	0.47	1.84	5.91	2.61	1.71
P	0.93	0.76	0.12	0.27	0.64

Table 2.1.33
Fitting the Hyperpoisson distribution to Turkish word lengths: Text 3

	Column						
Length	1	2	3	4	5	6	7
1	10.88	6.92	7.70	7.89	3.45	5.85	6.61
2	13.79	13.58	14.99	13.97	16.38	12.09	13.15
3	11.68	13.48	13.46	12.75	15.34	12.45	12.37
4	7.43	8.95	7.85	7.84	7.97	8.54	7.61
5	3.78	4.46	3.39	3.64	2.86	4.39	3.48
6	2.44	2.60	1.60	1.92	0.79	1.81	1.27
7					0.21	0.86	0.51
a	2.5517	2.0073	1.6661	1.8847	1.1661	2.0550	1.7829
b	2.0125	1.0231	0.8557	1.0648	0.2455	0.9950	0.8958
DF	3	3	3	3	2	3	3
X^2	6.46	3.58	1.87	6.21	2.07	6.38	1.29
P	0.09	0.31	0.60	0.10	0.36	0.09	0.73

	Column							
Rank	8	9	10	11	12	13	14	15
1	5.87	6.53	5.44	7.12	3.28	8.04	8.85	0.82
2	14.67	13.51	12.29	10.17	9.92	11.17	8.75	10.70
3	11.60	10.86	10.89	8.65	9.99	8.00	6.28	9.83
4	5.45	5.42	5.99	5.24	6.03	3.86	3.53	4.68
5	2.41	1.96	3.39	2.46	2.60	1.93	2.59	1.96
6		0.72		1.37	1.18			
a	1.1568	1.3141	1.4559	2.1011	1.5098	1.4763	2.6134	0.9887
b	0.4627	0.6347	0.6439	1.4707	0.4989	1.0625	2.6424	0.0761
DF	2	2	2	3	3	2	2	1
X^2	0.55	2.23	2.70	2.48	0.96	2.01	1.01	0.26
P	0.76	0.33	0.26	0.48	0.81	0.37	0.60	0.61

Table 2.1.34
Fitting the Hyperpoisson distribution to Turkish word lengths: Text 4

	Column						
Length	1	2	3	4	5	6	7
1	9.60	9.70	9.99	8.58	6.53	10.79	9.73
2	12.22	14.55	14.94	16.02	11.14	12.75	13.87
3	11.24	12.58	12.61	13.60	11.74	10.78	11.63
4	8.09	7.64	7.41	7.47	8.95	7.09	6.90
5	4.79	3.57	3.34	4.34	5.35	6.59	4.87

6 7 8	4.05	1.36 0.44 0.16	1.72		2.62 1.66		
a	3.3144	2.0402	1.9342	1.5561	2.7545	2.9673	2.0349
b	2.6041	1.3601	1.2926	0.8333	1.6133	2.5108	1.4273
DF	3	3	3	2	4	2	2
X^2	4.62	0.56	1.61	2.91	10.46	0.21	5.15
P	0.20	0.91	0.66	0.23	0.03	0.90	0.08

	Column							
Rank	8	9	10	11	12	13	14	15
1	9.97	7.42	11.16	6.84	0.54	4.58	4.01	4.49
2	14.66	14.07	13.20	11.02	9.20	7.94	6.12	6.98
3	11.53	12.02	9.03	10.00	11.91	7.75	6.02	5.89
4	6.19	6.63	4.35	6.32	8.01	5.26	4.38	3.40
5	2.52	3.86	2.26	3.06	5.33	2.74	2.52	2.24
6	1.12			1.75		1.74	1.94	
a	1.6908	1.5545	1.6239	2.0783	1.4005	2.2330	2.7781	1.8393
b	1.1493	0.8199	1.3729	1.2900	0.0818	1.2877	1.8219	1.1824
DF	3	2	2	3	1	3	3	2
X^2	2.11	1.67	2.14	2.58	0.22	1.55	2.61	1.20
P	0.55	0.43	0.34	0.46	0.64	0.67	0.46	0.55

Table 2.1.35
Fitting the Hyperpoisson distribution to Turkish word lengths: Text 5

	Column						
Length	1	2	3	4	5	6	7
1	10.97	12.38	0.43	5.42	6.53	7.41	7.21
2	14.40	15.71	18.06	14.23	11.14	12.97	12.27
3	11.94	11.80	18.14	14.59	11.74	11.89	11.45
4	7.24	6.30	9.22	9.29	8.95	7.39	7.37
5	3.46	2.61	3.14	4.29	5.35	3.47	3.61
6	2.00	1.21	1.00	2.17	2.62	1.31	2.09
7					1.66	0.56	
a	2.2544	1.8423	1.0292	1.6824	2.7545	1.9263	2.0678
b	1.7185	1.4519	0.0245	0.6412	1.6133	1.1007	1.2151
DF	3	3	1	3	4	3	3
X^2	2.11	0.76	0.61	2.12	10.46	4.37	4.31
P	0.55	0.86	0.43	0.55	0.03	0.22	0.23

	Column							
Rank	8	9	10	11	12	13	14	15
1	9.47	6.56	7.82	6.83	2.58	0.75	4.41	2.42
2	8.92	10.87	9.47	7.62	9.76	12.35	5.59	5.02
3	7.61	10.27	8.29	7.00	9.78	10.13	5.27	5.79
4	5.94	6.77	5.69	5.45	5.65	4.26	3.95	4.63
5	4.28	3.43	3.21	3.69	2.29	1.51	2.46	2.84
6	2.85	2.10	1.54	2.21	0.95		1.31	1.41
7	3.93		0.98	1.19			1.00	0.89
				1.00				
a	9.1631	2.1902	3.1681	5.1714	1.3637	0.8629	3.6718	2.6077
b	9.7341	1.3201	2.6174	4.6344	0.3612	0.0522	2.8965	1.2590
DF	4	3	3	4	2	1	3	3
X^2	6.54	9.49	0.61	2.38	10.51	0.30	1.75	4.08
P	0.16	0.02	0.89	0.67	0.01	0.58	0.63	0.25

In 5 texts of Turkish data there are 8 columns which cannot be fitted satisfactorily by the Hyperpoisson distribution. There are, of course other, simple, very adequate distributions but here we shall not show them separately because texts/columns of that kind should be analyzed literally/qualitatively and in each column separately the "cause" of deviation (boundary conditions) should be found.

In **French** there were some strongly concentrated columns containing only two or three types of length. Fitting a distribution with two parameters would not yield any chi-square value because there arise zero degrees of freedom. Hence we applied rather special cases of the Hyperpoisson, namely the Poisson distribution, obtained in the case $b = 1$, and the geometric distribution which is a limiting case for $a \to \infty$, $b \to \infty$ and $a/b \to q$, i.e.

Poisson d.: (2.6) $$P_x = \frac{a^x e^{-a}}{x!},$$

Geometric d.: (2.7) $$P_x = pq^x$$

Both distributions are shifted one step to the right. But there is a column in Text 1 (column 13) where there are only two length classes. In this case one can add a third class with frequency 0; by means of such a "modification" one can obtain e.g. the Prasad distribution (cf. Wimmer, Altmann 1999) below.

The results show that French is on a way towards monosyllabism, though in script one can see several vowels in a word.

The values are presented in Tables 2.1.36 to 2.1.40 and the fitting in Tables 2.1.41 to 2.1.45.

Table 2.1.36
French lengths: Text 1

	Columns									
Length	1	2	3	4	5	6	7	8	9	10
1	36	28	37	31	29	26	28	30	21	29
2	10	19	7	10	12	11	10	6	8	8
3	4	3	4	5	4	8	4	6	10	1
4			2	4	3	2	1	1	1	3
5									1	
Sum	50	50	50	50	48	47	43	43	41	41
WCS	.72	.56	.74	.62	.60	.55	.65	.70	.51	.71

	Columns									
Length	11	12	13	14	15	16	17	18	19	20
1	23	21	29	17	18	15	19	13	18	11
2	8	8	5	7	5	5	3	5	2	6
3	3	2		5	2	4	2	4	2	4
4	3	1		2	1	0		1	1	
5	1	2				1				
Sum	38	34	34	31	26	25	24	23	23	21
WCS	.61	.62	.85	.55	.69	.60	.79	.56	.78	.52

Table 2.1.37
French lengths: Text 2

	Columns									
Length	1	2	3	4	5	6	7	8	9	10
1	38	25	27	30	32	26	28	30	24	25
2	10	15	14	14	10	17	12	10	10	11
3	0	8	5	5	6	5	4	6	8	8
4	2	1	3		1	1	2	1	3	2
5		1					1		1	
Sum	50	50	49	49	49	49	47	47	46	46
WCS	.76	.50	.55	.61	.65	.53	.60	.64	.52	.54

	Columns									
Length	11	12	13	14	15	16	17	18	19	20
1	28	25	24	23	24	24	23	15	17	10
2	9	9	12	11	7	5	7	11	7	12
3	7	10	6	5	1	7	5	5	7	6

4 5 6	1	1		3	4 0 1	1	2	2 1	1 1	2 1
Sum	45	45	42	42	37	37	37	34	33	31
WCS	.62	.56	.57	.55	.65	.65	.62	.44	.52	.32

Table 2.1.38
French lengths: Text 3

	Columns									
Length	1	2	3	4	5	6	7	8	9	10
1 2 3 4 5 6	39 6 1 4	33 10 5 1 1	34 10 4 1 1	32 10 4 2 1	29 11 4 4 1	29 15 3 1 1	33 10 4 1	27 11 9 1	25 12 6 2 1	25 11 5 2 0 1
Sum	50	50	50	49	49	49	48	48	46	44
WCS	.78	.66	.68	.65	.59	.59	.69	.56	.54	.57

	Columns									
Length	11	12	13	14	15	16	17	18	19	20
1 2 3 4 5	24 8 9 0 2	25 8 6 2	27 8 4 1	24 8 5 1	19 6 6 3	23 5 4 0 1	21 7 3 1	24 5 2 1	21 5 1 3	20 7 1 2
Sum	43	41	40	38	34	33	32	32	30	30
WCS	.56	.61	.68	.63	.56	.70	.66	.75	.70	.67

Table 2.1.39
French lengths: Text 4

	Columns									
Length	1	2	3	4	5	6	7	8	9	10
1 2 3 4 5 6	36 12 1 1	25 18 6	28 11 8 2	29 12 6 1	27 15 4 0 1	25 15 4 1 1	27 13 5 0 0 1	26 11 4 2 1	24 10 9 0 0 1	26 6 8 1 2
Sum	50	49	49	48	47	46	46	44	44	43
WCS	.72	.51	.57	.60	.57	.54	.59	.59	.55	.60

	Columns									
Length	11	12	13	14	15	16	17	18	19	20
1	23	25	19	20	27	17	16	16	12	19
2	14	9	10	12	5	10	13	8	9	4
3	6	6	6	4	1	5	2	4	6	2
4		1	2	2	5	3	1	2	1	2
5			2	1		0	1			
						0				
						1				
Sum	43	41	39	39	38	36	33	30	28	27
WCS	.53	.61	.49	.51	.71	.47	.48	.53	.43	.70

Table 2.1.40
French lengths: Text 5

	Columns									
Length	1	2	3	4	5	6	7	8	9	10
1	40	29	25	26	26	27	25	25	30	23
2	8	13	10	15	12	8	11	8	8	13
3	2	3	9	4	7	5	6	7	5	2
4		5	5	5	4	8	2	6	2	4
5			1			1	2		1	1
Sum	50	50	50	50	49	49	46	46	46	43
WCS	.80	.58	.50	.52	.53	.55	.54	.54	,65	.53

	Columns									
Length	11	12	13	14	15	16	17	18	19	20
1	26	22	30	22	19	18	15	14	15	14
2	6	13	4	10	6	8	9	9	6	5
3	4	5	0	3	9	3	5	4	3	3
4	6	1	4	2		2	2	2	3	4
5								1	1	
Sum	42	41	38	37	34	31	31	30	28	26
WCS	.62	.54	.79	.59	.56	.58	.48	.47	.54	.54

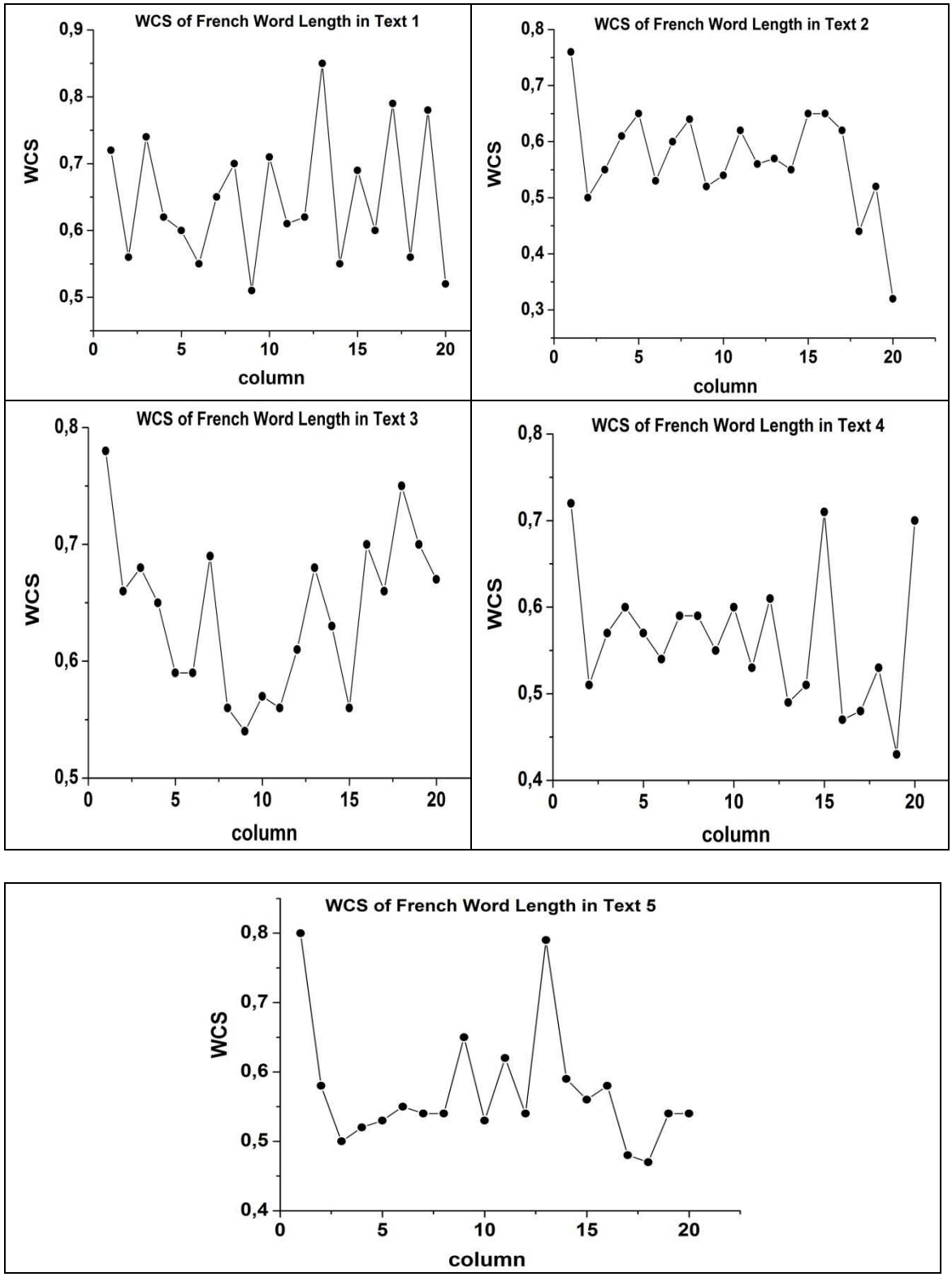


Figure 2.1.6. WCS of French word lengths

Table 2.1.41
Fitting the geometric distribution to French length data: Text 1
(P = Poisson d., PP = positive Poisson d., SP = Singh-Poisson d.)

	Column						
Length	1	2 P	3	4	5	6	7
1	35.96	29.53	34.47	29.42	29.16	26.25	28.41
2	10.10	15.55	10.71	12.11	11.45	11.59	9.64
3	3.94	4.92	3.32	4.99	4.49	5.12	3.27
4			1.50	3.49	2.90	4.05	1.68
p	0.7192	a=0.5268	0.6894	0.5883	0.6075	0.5585	0.6608
DF	1	1	2	2	2	2	2
X^2	0.02	1.59	1.77	0.53	0.08	2.69	0.46
P	0.97	0.21	0.41	0.77	0.96	0.26	0.80

	Column							
Length	8	9	10	11	12	13Pras	14	15
1	27.67	20.43	29.11	21.46	20.05	28.16	17.06	17.58
2	9.86	10.25	8.44	9.34	8.23	3.42	7.67	5.69
3	3.52	5.14	2.45	4.07	3.38	2.42	3.45	1.85
4	1.95	2.58	1.00	1.77	1.39		2.82	0.88
5		2.60		1.37	0.96			
p	0.6435	0.4984	0.7100	0.5646	0.5896	a=0.415	0.5504	0.6760
DF	2	3	1	3	2	1	2	1
X^2	3.93	7.05	0.11	1.53	0.79	3.17	0.99	0.12
P	0.14	0.07	0.74	0.67	0.67	0.07	0.61	0.73

	Column				
Length	16	17 PP	18	19 SP	20
1	14.58	17.44	12.89	17.95	11.49
2	6.08	5.30	5.67	2.29	5.20
3	2.53	1.26	2.49	1.54	4.31
4	1.06		1.95	1.23	
5	0.76				
p	0.5831	a=0.6079	0.5605	a = 2.0129;α = 0.3187	0.5472
DF	2	1	2	1	1
X^2	1.42	1.57	1.46	0.22	0.16
P	0.49	0.21	0.48	0.64	0.68

Table 2.1.42
Fitting the geometric distribution to French length data: Text 2
(PP = positive Poisson d.)

	Column						
Length	1	2	3	4	5	6	7
1	38.59	27.30	28.45	31.48	31.41	29.17	28.38
2	8.80	12.39	11.93	11.26	11.28	11.81	11.24
3	2.01	5.63	5.00	6.26	4.05	4.78	4.45
4	0.59	2.55	3.62		2.27	3.25	1.76
		2.12					1.16
p	0.7719	0.5461	0.5805	0.6425	0.6410	0.5952	0.6039
DF	1	3	2	1	2	2	3
X^2	0.31	3.28	0.54	0.99	1.80	4.20	0.16
P	0.58	0.35	0.76	0.32	0.41	0.12	0.98

	Column						
Length	8	9	10	11	12PP	13	14
1	29.66	23.99	25.40	27.16	22.06	25.13	23.69
2	10.94	11.48	11.38	10.77	14.22	10.09	10.33
3	4.04	5.49	5.09	4.27	6.11	6.78	4.50
4	2.36	2.63	4.13	2.80	2.61		3.48
5		2.41					
6							
p	0.6310	0.5216	0.5521	0.6035	a=1.2893	0.5983	0.5642
DF	2	3	2	2	2	1	2
X^2	1.82	2.21	2.78	3.22	5.78	0.50	0.18
P	0.40	0.53	0.25	0.20	0.06	0.48	0.91

	Column					
Length	15	16	17	18	19	20
1	22.00	21.62	22.14	17.31	16.89	14.21
2	8.92	8.99	8.89	8.50	8.25	7.70
3	3.62	3.73	3.57	4.17	4.03	4.17
4	1.47	2.66	2.40	2.05	1.97	2.26
5	0.59			1.98	1.88	2.67
6	0.41					
p	0.5945	0.5844	0.5984	0.5090	0.5117	0.4584
DF	2	2	2	3	3	3
X^2	5.09	5.92	1.07	1.69	3.27	5.53
P	0.08	0.05	0.58	0.64	0.35	0.14

In order to fit column 1 in Text 3 we have chosen the Prasad distribution which is a mixture of the 1-displaced geometric distribution and the beta distribution with parameters (2,a). It appears that this way of expressing boundary conditions is a legitimate mathematical approach: one of the parameters is not constant. From the linguistic point of view, it can be found only after many texts and languages have been studied. The Prasad distribution has only one parameter and is defined as

$$(2.8) \quad P_x = \frac{2a(a+1)}{(a+x-1)(a+x)(a+x+1)}, \; x=1,2,3,...$$

The parameter *p* of the geometric distribution is considered itself a variable distributed according to beta distribution (cf. Wimmer, Altmann 1999)

Table 2.1.43
Fitting the geometric distribution to French length data: Text 3
(Pd = Prasad d.)

	Column						
Length	1Pd	2	3	4	5	6	7
1	38.16	32.28	33.30	30.81	27.60	30.87	32.99
2	6.54	11.44	11.12	11.44	12.05	11.42	10.32
3	2.29	4.05	3.72	4.25	5.26	4.23	3.23
4	3.01	1.44	1.24	1.58	2.30	1.56	1.47
5		0.79	0.62	0.93	1.78	0.92	
p	a = 0.6206	0.6456	0.6660	0.6288	0.5633	0.6300	0.6872
DF	2	2	2	2	3	2	2
X^2	1.12	0.44	0.16	0.34	2.07	1.68	0.34
P	0.57	0.80	0.92	0.84	0.56	0.43	0.84

	Column						
Length	8	9	10	11	12	13	14
1	26.76	26.01	25.48	22.29	24.19	26.67	23.68
2	11.84	11.30	10.73	10.74	9.92	8.89	8.92
3	5.24	4.91	4.52	5.17	4.07	2.96	3.36
4	4.16	2.14	1.90	2.49	2.83	1.48	2.03
5		1.64	0.80	2.32			
6			0.58				
p	0.5576	0.5654	0.5790	0.5183	0.5899	0.6667	0.6231
DF	2	3	3	3	2	2	2
X^2	5.16	0.58	0.18	6.20	1.56	0.61	1.42
P	0.08	0.90	0.98	0.10	0.46	0.74	0.49

	Column					
Length	15	16	17	18	19	20
1	17.95	21.59	20.99	23.12	20.35	20.35
2	8.47	7.46	7.22	6.42	6.55	6.55
3	4.00	2.58	2.48	1.78	2.11	2.11
4	3.58	0.89	1.30	0.68	1.00	1.00
5		0.47				
p	0.5279	0.6543	0.6560	0.7225	0.6782	0.6782
DF	2	2	2	1	1	1
X^2	1.88	1.78	0.18	0.46	0.64	0.64
P	0.39	0.41	0.91	0.50	0.42	0.42

Table 2.1.44
Fitting the geometric distribution to French length data: Text 4

	Column						
Length	1	2	3	4	5	6	7
1	37.13	28.47	27.96	29.65	29.42	27.32	28.55
2	9.56	11.93	12.01	11.33	11.00	11.09	10.83
3	2.46	8.60	5.16	4.33	4.12	4.50	4.11
4	0.85		3.88	2.68	1.54	1.83	1.56
					0.92	1.25	0.59
							0.36
p	0.7426	0.5810	0.5706	0.6177	0.6259	0.5940	0.6206
DF	1	1	2	2	2	3	2
X^2	1.18	4.30	2.56	1.75	2.52	2.06	1.62
P	0.28	0.04	0.28	0.42	0.28	0.56	0.44

	Column						
Length	8	9	10	11	12	13	14
1	26.25	23.56	23.21	25.19	24.79	19.75	21.69
2	10.59	10.94	10.68	10.43	9.80	9.75	9.63
3	4.27	5.08	4.92	7.38	3.88	4.81	4.27
4	1.72	2.36	2.26		2.54	2.37	1.90
5	1.17	1.10	1.93			2.31	1.52
6		0.95					
p	0.5965	0.5355	0.5398	0.5857	0.6045	0.5064	0.5560
DF	3	3	3	1	2	3	3
X^2	0.10	6.00	5.03	1.67	2.16	0.43	0.91
P	0.99	0.11	0.17	0.20	0.34	0.93	0.82

	Column					
Length	15	16	17	18	19	20
1	26.70	18.42	18.55	16.73	13.85	18.00
2	7.94	9.00	8.12	7.40	7.00	6.00
3	2.36	4.39	3.56	3.27	3.54	2.00
4	1.00	2.14	1.56	2.60	3.61	1.00
5		1.05	1.21			
6		0.51				
7		0.49				
p	0.7026	0.5117	0.5622	0.5575	0.4948	0.6667
DF	1	3	3	2	2	1
X^2	3.16	1.18	4.20	0.38	4.42	1.06
P	0.08	0.76	0.24	0.83	0.11	0.30

Table 2.1.45
Fitting the geometric distribution to French length data: Text 5

	Column					
Length	1	2	3	4	5	6SP
1	40.02	28.35	24.28	27.02	26.44	25.75
2	7.99	12.28	12.49	12.42	12.17	7.73
3	1.99	5.31	6.42	5.71	5.61	7.37
4		4.06	3.30	4.85	4.78	4.69
5			3.50			3.45
p	0.8004	0.5671	0.4857	0.5405	0.5395	a=1.9067. α=0.5572
DF	1	2	3	2	2	2
X^2	0.0001	1.28	4.21	1.09	0.49	4.92
P	0.99	0.53	0.24	0.58	0.78	0.09

		Column						
Length	7	8	9	10	11	12	13	14
1	24.93	22.98	28.34	23.05	22.02	24.10	27.58	22.67
2	11.42	11.50	10.88	10.69	10.48	9.93	7.56	8.78
3	5.23	5.76	4.18	4.96	4.98	4.09	2.07	3.40
4	2.40	5.77	1.60	2.30	4.52	2.87	0.78	2.15
5	2.03		1.00	1.99				
p	0.5419	0.4995	0.6161	0.5360	0.5242	0.5878	0.7258	0.6128
DF	3	2	2	3	2	2	1	2
X^2	0.19	1.52	1.08	4.01	3.31	2.55	2.35	0.25
P	0.98	0.47	0.58	0.26	0.19	0.28	0.13	0.88

	Column					
Length	15	16	17	18	19	20
1	17.32	18.34	16.41	15.36	14.10	12.69
2	8.50	7.49	7.72	7.50	7.00	6.50
3	8.18	3.06	3.64	3.66	3.47	3.33
4		2.11	3.23	1.78	1.72	3.49
5				1.70	1.70	
p	0.5095	0.5916	0.5293	0.5121	0.5036	0.4880
DF	1	2	2	3	3	2
X^2	0.98	0.05	1.31	0.77	1.50	0.59
P	0.32	0.98	0.52	0.86	0.68	0.75

We want to venture the conjecture that the word length frequencies inside of a column of (1.1) follow the Hyperpoisson distribution but boundary conditions may lead to simpler models. Further, we may assume that some of the functions of the distributions are correlated with the WCS. However, in order to test this it would be necessary to compute all known characteristics of the individual distributions, e.g. moments, their functions, entropy, repeat rate, Gini's coefficient, etc.

2.2. Polysemy

Polysemy of a word can be stated on the basis of dictionaries, some of which are accessible via the Internet. The sequence corresponding to a sentence, containing the number of meanings, can easily be processed.

The text used here has been extracted from Persian journalistic texts, mainly from Hamshahri Newspaper1[1]

Here, by polysemy we mean the number of different meanings of a word. For counting these numbers for Persian words, we made use of the version 4.1 of the Moeen Dictionary of Persian, the most recent version available as software. In the tables below, it can be seen that there also exists a vertical structure but now the problem is slightly different. If a word belongs to more than one part-of-speech, it may have a special meaning given by the context, but the dictionary lists all possible meanings that it can have. In that case one may obtain distributions with two or more local maxima (cf. columns 1 and 5 in Table 3.5.1), and one may apply a generalized, modified or mixed distribution consisting of two components. The situation is here more complex. We shall try to confine ourselves to the Poisson family, using the most adequate special case for fitting. The causes for the boundary conditions, i.e. exceptions from the regularity, cannot be found but if one performs the analysis of several texts in various text types

and languages, one will be able to predict the adequacy of a special distribution. Here we shall use the following distributions:

$$\text{Mixed Poisson d.: } P(X) = \frac{\alpha a^x e^{-a}}{x!} + (1-\alpha)\frac{b^x e^{-b}}{x!},\ x = 0,1,2,...$$

which should be displaced one step to the right, and

$$\text{Singh-Poisson d.: } P(X) = \begin{cases} 1-\alpha+\alpha e^{-a}, & x = 1 \\ \dfrac{\alpha a^{x-1} e^{-a}}{(x-1)!}, & x = 2,3,... \end{cases}$$

which is already given in displaced form. Of course, other distributions might be adjusted but preliminarily we shall use only the above mentioned ones. The results of fitting are presented in Table 2.2.2. In column 1 the pooling criterion for theoretical classes for the chi-square test was 5, (i.e. the smallest classes have been pooled until the sum 5 was reached), in all other classes no pooling was necessary, i.e. the original classes have not been changed.

Table 2.2.1
Polysemy in Persian text

Poly-	Column										
semy	1	2	3	4	5	6	7	8	9	10	11
1	27	24	27	23	16	33	28	20	19	26	22
2	5	13	11	11	6	7	9	8	8	6	12
3	1	5	5	5	8	3	2	9	7	4	5
4	4	2	3	5	7	2	2	7	6	3	4
5	10	3	2	4	10	3	4	2	2	2	4
6	1	1	1	9	1	0	1	1	0	3	0
7	1	1	1	1	0	0	2	2	3	1	0
8	0	0		0	0	2	0	0	1	0	0
9	1	1		1	2		1	1	4	3	0
10							0				0
11							0				1
12							1				
SUM	50	50	50	50	50	50	50	50	50	48	48
WCS	.54	.48	.54	.46	.32	.66	.56	.40	.38	.54	.46

Poly-semy	Column										
	12	13	14	15	16	17	18	19	20	21	22
1	16	18	20	16	17	13	13	14	16	11	7
2	11	12	8	8	4	7	6	6	4	4	6
3	10	10	3	9	8	9	6	5	4	3	5
4	3	5	7	2	2	0	2	3	2	4	1
5	3	1	5	1	2	6	2	1	2	2	2
6	2	0	0	1	0	1	1	0	0		0
7	1	1	2	1	3	0	2	0	1		0
8	0	0		0	0	0	0	0	0		0
9	2	0		1	2	2	1	2	1		2
10		1		0							
11				0							
12				0							
13				0							
14				0							
15				0							
16				1							
SUM	48	48	45	40	38	38	33	31	30	24	23
WCS	.33	.38	.44	.40	.45	.34	.39	.45	.53	.46	.30

For example, in column 1 of Table 2.2.1 there are 27 words with one meaning, 5 words with two meanings, 1 word with three meanings etc.

Table 2.2.2
Fitting the mixed Poisson and the Singh-Poisson distributions to Persian vertical polysemy distributions (SIPO = Singh-Poisson, MixPO = Mixed Poisson)

Poly-semy	Column							
	1 SIPO	2 MixPO	3 MixPO	4 MixPO	5 SIPO	6 MixPO	7 MixPO	8 SIPO
1	26.98	25.13	26.91	21.28	15.29	34.45	29.39	8.99
2	2.91	11.96	11.10	9.91	6.01	7.06	8.52	8.10
3	4.80	4.15	4.83	5.75	8.61	2.52	2.39	9.20
4	5.28	2.56	3.21	5.87	8.22	2.18	2.06	6.97
5	4.36	2.20	2.07	5.66	5.89	1.71	2.21	3.96
6	2.88	1.71	1.10	4.47	3.37	1.09	1.98	1.80
7	1.58	1.12	0.77	2.96	1.61	0.58	1.49	0.68
8	0.75	0.63		1.67	0.66	0.42	0.95	0.22
9	0.47	0.53		1.42	0.34		0.54	0.08
10							0.27	
11							0.12	
12							0.08	
a	3.2999	3.9439	2.6830	3.9642	2.8653	3.1850	4.4933	2.2721

b	-	0.4462	0.3291	0.3731	-	0.1700	0.2712	-
α	0.4781	0.2214	0.2785	0.4896	0.7362	0.1929	0.2325	0.6914
DF	1	4	2	5	4	2	4	3
X^2	3.62	1.06	0.03	8.64	4.94	1.14	2.27	1.56
P	0.06	0.90	0.98	0.12	0.29	0.57	0.69	0.67
Poly-semy	Column							
	9 MixPO	10 MixPO	11 MixPO	12 MixPO	13 SIPO	14 SIPO	15 MixPO	16 SIPO
1	15.51	26.09	21.75	14.65	17.99	19.15	13.72	19.15
2	12.04	6.77	12.22	14.11	12.15	6.29	13.25	3.93
3	5.60	2.72	5.23	7.59	9.72	7.47	6.43	5.11
4	3.30	2.95	3.28	3.86	5.18	5.92	2.18	4.43
5	3.10	3.05	2.41	2.57	2.07	3.52	0.74	2.88
6	3.06	2.56	1.57	1.98	0.66	1.67	0.47	1.50
7	2.65	1.79	0.87	1.42	0.18	0.98	0.51	0.65
8	1.97	1.07	0.41	0.90	0.04		0.57	0.24
9	2.75	1.00	0.17	0.93	0.01		0.56	0.11
10			0.06		0.002		0.49	
11			0.03				0.39	
12							0.28	
13							0.18	
14							0.11	
15							0.06	
16							0.06	
a	5.2279	4.1978	3.3271	4.4647	1.5996	2.3766	7.9061	2.6009
b	0.7494	0.2235	0.5017	0.9324	-	-	0.9652	-
α	0.3475	0.3266	0.2674	0.2315	0.7835	0.6333	0.0999	0.5359
DF	5	4	3	4	2	3	4	3
X^2	9.20	1.89	2.99	1.98	0.02	4.16	4.03	5.98
P	0.10	0.76	0.39	0.74	0.99	0.24	0.40	0.11

Poly-semy	Column					
	17 SIPO	18 MixPO	19 SIPO	20 SIPO	21 SIPO	22 MixPO
1	13.20	11.08	17.09	14.67	10.83	7.18
2	4.64	8.10	4.05	4.17	3.82	7.44
3	6.39	3.84	4.27	4.61	4.04	3.96
4	5.87	2.46	3.00	3.40	2.85	1.61
5	4.04	2.21	1.59	1.88	2.46	0.77
6	2.23	1.93	0.67	0.83		0.56
7	1.02	1.45	0.24	0.31		0.48
8	0.40	0.94	0.07	0.10		0.38
9	0.20	1.00	0.02	0.04		0.62

a	2.7548	4.5374	2.1108	2.2128	2.1146	5.6691
b	-	0.6899	-	-	-	1.0287
α	0.6969	0.3378	0.5104	0.5737	0.6241	0.1273
DF	4	4	2	3	2	2
X^2	9.85	2.71	1.90	1.21	0.83	1.31
P	0.04	0.61	0.43	0.75	0.66	0.52

The WCS of Persian polysemy is displayed in Figure 2.2.1

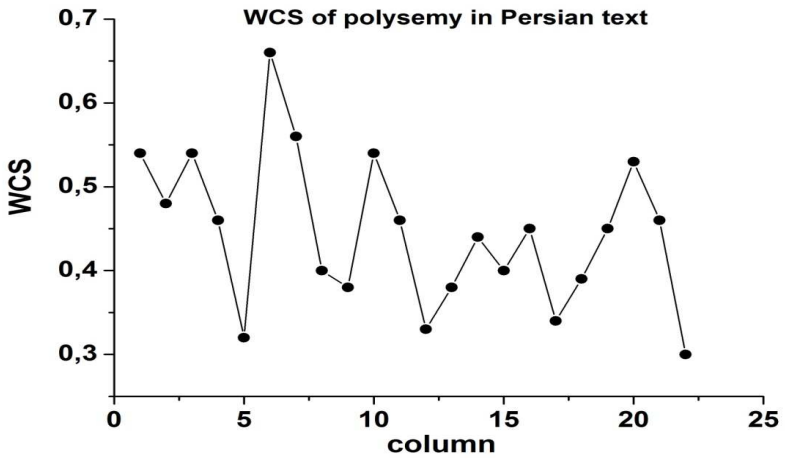


Figure 2.2.1. WCS of polysemy in Persian texts

2.3. Frequency strings

The given text is analyzed for word frequencies having two variants: either one considers lemmas or all forms of a word. Then each word/lemma in the text will be substituted by its frequency. The frequency can be taken either from a frequency dictionary of the language or from the given text. One obtains a sequence of numbers. Each sentence is then a string of integers which can be further evaluated. It is advisable to consider rather lemmas, because a strongly synthetic language could display very low frequencies if the number of forms is very large.

There are still other problems which may be solved ad hoc before one begins to count, e.g. the status of compounds e.g. *fall velocity*, full personal or geographic names, compounds like *The United States of America*, etc. Problems of this kind must be solved according to commonly accepted lexicological conventions.

This aspect is not in agreement with Skinner's hypothesis because frequent words are mostly synsemantics, auxiliaries, etc., hence their distances are usually rather short. As to autosemantics, one can expect many short distances and a decreasing number of longer distances – in agreement with the Skinner

hypothesis. Replacing the words by frequencies, one obtains a string with high variations whose underlying laws must be carefully studied.

A third possibility may be tested in the following way: Replace all synsemantics by a zero and consider only autosemantics at the given place and evaluate the frequencies and the weighted consensus strings in this new matrix.

Here we shall present explicitly the positioning of frequencies in the German poem *Der Erlkönig* by J.W.v. Goethe. The basic data are presented in the Appendix.

First we consider the rank-frequency distributions in the individual columns. The basic data are displayed in Table 2.3.1. The columns 8 and 9 contain only small numbers not appropriate for testing, and will be omitted. The results of fitting the Zipf-distribution are presented in Table 2.3.2.

Table 2.3.1
Rank-frequency distribution in *Der Erlkönig* by J.W.v. Goethe

	Column						
Rank	1	2	3	4	5	6	7
1	8	8	8	11	13	10	8
2	8	8	8	6	5	5	6
3	5	5	5	4	4	3	2
4	4	4	4	4	3	3	1
5	2	3	2	3	3	3	1
6	2	2	2	2	1	2	1
7	2	2	1	1	1	2	
8	1		1	1	1	1	
9			1		1	1	
Sum	32	32	32	32	32	30	29

Table 2.3.2
Fitting the zeta distribution to the data in Table 2.3.1

	Column						
Rank	1	2	3	4	5	6	7
1	9.94	9.60	10.39	11.31	12.72	9.75	8.95
2	5.69	5.92	5.55	5.85	5.77	5.20	3.83
3	4.11	4.46	3.84	3.98	3.63	3.60	2.33
4	3.26	3.65	2.96	3.02	2.61	2.78	1.64
5	2.72	3.13	2.42	2.45	2.03	2.27	1.25
6	2.35	2.75	2.05	2.06	1.65	1.92	1.00
7	2.08	2.47	1.78	1.78	1.36	1.67	
8	1.86		1.58	1.56	1.18	1.48	
9			1.42		1.04	1.33	
a	0.8047	0.6972	0.9051	0.9510	1.1416	0.9063	1.2244

R	8	7	9	8	9	9	6
DF	5	4	6	5	6	6	2
X^2	2.33	1.39	3.10	1	1.06	0.68	1.65
P	0.80	0.85	0.80	0.96	0.98	0.99	0.44

Though we have to do here with a poem, the fitting is excellent.

The WCS is given as

$$WCS = [8/32, 8/32, 8/32, 11/32, 13/32, 10/32, 8/30, 2/29] =$$
$$= [0.25, 0.25, 0.25, 0.34, 0.41, 0.31, 0.27, 0.07]$$

and presented in Figure 2.3.1. The text yields a smooth series of frequencies that could be described by a continuous function.

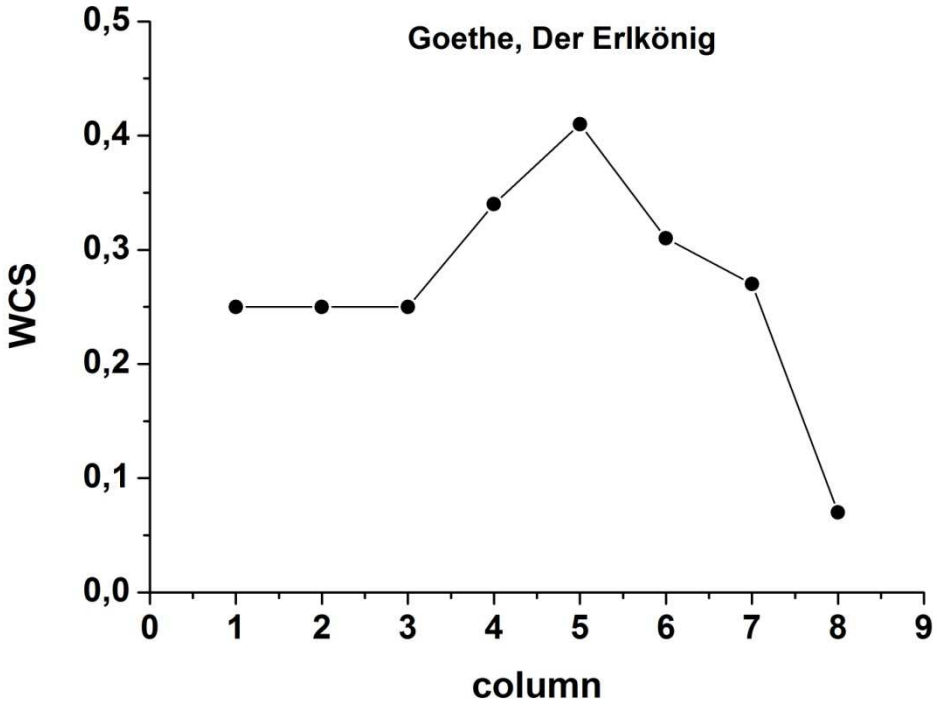


Figure 2.3.1. WCS for frequency positioning in Erlkönig.

2.4. Parts of speech

The symbolization of parts of speech may be performed according to authoritative grammars. A word can belong simultaneously to different classes, and in that case one must decide whether *the book* is Article+N/V, or simply N as a whole, because there is also the combination *to book* where one has either Prep+N/V or simply V. One may, of course, adhere to the function of the word in the phrase or sentence and obtain Article + N, and Prep + V. In many languages there are similar problems, e.g. in German, many adjectives are at the same time

adverbs, and one can distinguish them only in the sentence; all verbs in infinitive are at the same time nouns, detachable prefixes may become adverbs, etc. Here we shall perform the analysis according to the function in sentence. One must be aware of the fact that the old Latin classification is not always sufficient, one should rather adhere to modern grammars.

For German we analyzed “Der Erlkönig”, a poem by J.W. Goethe, and obtained verse-wise the results presented in Table 2.4.1.

Table 2.4.1
Parts of speech in verse-position in Erlkönig

	Column/position								
POS	1	2	3	4	5	6	7	8	9
N	2	14	7	8	6	7	4	3	2
V	4	10	5	7	3	8	2	-	-
P	19	1	12	5	5	3	8	3	-
Pr	2	-	3	1	3	3	-	-	-
Art	2	1	3	4	3	1	1	-	-
C	2	-	1	-	5	1	1	-	-
A	-	6	-	3	3	1	3	-	-
Av	1	-	1	3	2	5	1	3	2
Part	-	-	-	1	-	1	-	2	-
Sum	32	32	32	32	32	32	20	11	4

As is usual with rank-frequency distributions – the only possible way with qualitative variables – we obtain the reordering in Table 2.4.2

Table 2.4.2
POS rank-frequency of sequences in verse-position in *Erlkönig*

	Column/position							
Rank	1	2	3	4	5	6	7	8
1	19	14	12	8	6	8	8	3
2	4	10	7	7	5	7	4	3
3	2	6	5	5	5	5	3	3
4	2	1	3	4	3	3	2	2
5	2	1	3	3	3	3	1	
6	2		1	3	3	1	1	
7	1		1	1	3	1	1	
8			1	2	1			
9					1			
Sum	32	32	32	32	32	32	20	11
WCS	.59	.44	.38	.25	.19	.25	.40	.28

The ”higher” positions (columns) contain very small numbers of elements hence fitting would not be efficient. Since we have to do with classes, we apply the simple right truncated zeta (Zipf) distribution according to the formula

$$(2.9) \quad P_x = \frac{x^{-a}}{F(R)}, \quad x = 1, 2, 3, ..., R$$

where $F(R) = \sum_{i=1}^{R} i^{-a}$ is the normalizing constant. For more extensive data one can apply the Zipf-Mandelbrot distribution. The results of fitting are displayed in Table 2.4.3.

Table 2.4.3
Fitting the zeta distribution to rank-frequencies of word classes in German

	Column							
Rank	1	2	3	4	5	6	7	8
1	17.00	15.21	12.72	9.32	6.69	9.38	8.23	3.15
2	6.01	6.90	6.18	5.59	4.69	5.30	3.87	2.79
3	3.27	4.34	4.06	4.15	3.81	3.80	2.49	2.60
4	2.13	3.13	3.01	3.36	3.28	3.00	1.82	2.47
5	1.52	2.42	2.38	2.85	2.93	2.49	1.43	
6	1.16		1.97	2.49	2.67	2.14	1.17	
7	0.92		1.68	2.22	2.47	1.89	0.99	
8				2.01	2.30			
9					2.17			
a	1.4998	1.1407	1.0401	0.7369	0.5129	0.8236	1.0875	0.1750
R	7	5	7	8	9	7	7	4
DF	3	2	4	5	6	4	3	1
X^2	1.97	4.41	1.28	2.13	1.32	2.26	0.27	0.17
P	0.58	0.11	0.86	0.83	0.97	0.69	0.98	0.68

As can be seen, the rank-frequency distributions of the columns follow the usual Zipf-distribution.

As weighted consensus string, we obtain

WCS(German) = (19/32, 14/32, 12/32, 8/32, 6/32, 7/30, 8/30, 3/11, 2/4).
= (0.594, 0.438, 0.375, 0.25, 0.189, 0.27, 0.4, 0.273, 0.5)

which begins, as expected, with rather greater values which diminish, and finally increase again.

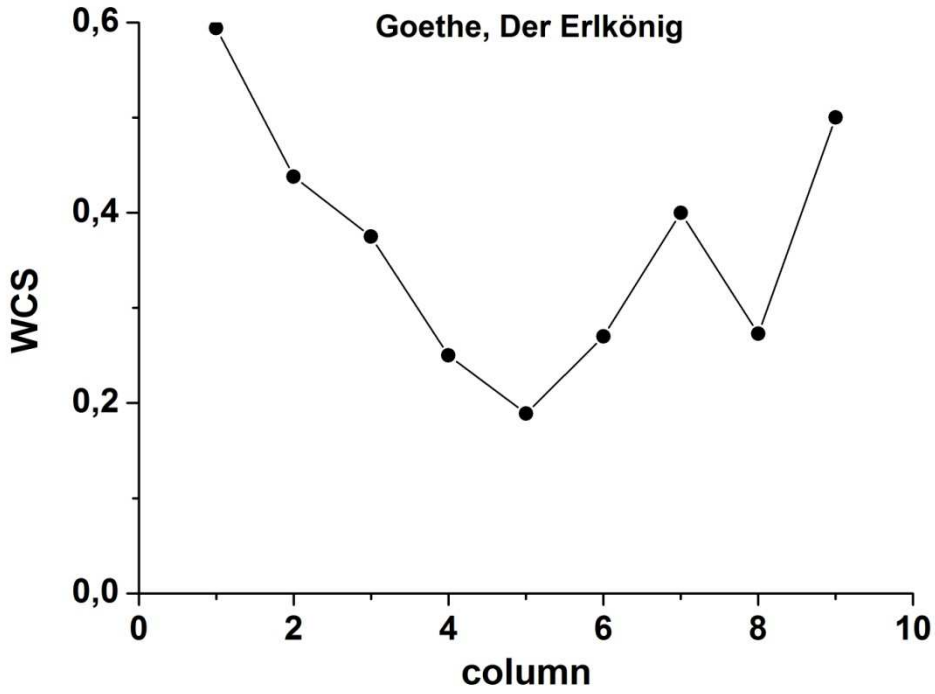


Figure 2.4.1, WCS for the *Erlkönig*

The smoother the weighted consensus string, the more homogeneous is the text, i.e. the more similar is the sentence/verse structure.

In Chinese, the different kinds of words are listed in the Chinese Lexical Analysis System developed by ICTCLAS (Institute of Computing Technology, Chinese Academy of Sciences), which is the same system for word segmentation in LCMC (second version) (Xiao 2012:164-166). One can obtain the analysis according to the Table 2.4.4.

Table 2.4.4
ICTCLAS Annotation System

POS	**Abr.**	**Meaning**
Noun	n	Noun
	nr	person's name
	ns	name of place
	nt	name of organization
	nz	other proper noun
	nl	fixed expression serving as noun
	ng	monosyllabic morpheme as noun
Verb	v	Verb
	vd	auxiliary verb

	vn	verb serving as noun
	vshi	Be structure
	vyou	have/there be structure
	vf	tendency verb
	vx	dummy verb
	vi	intransitive verb
	vl	fixed expression serving as verb
	vg	monosyllabic morpheme as verb
Adjective	a	
	ad	adjective serving as verb
	an	adjective serving as noun
	ag	monosyllabic morpheme as adjective
	al	fixed expression serving as adjective
Adverb	d	Adverb
	dg	monosyllabic morpheme as adverb
Differentiating Word	b	differentiating word
	bl	monosyllabic morpheme as differentiating word
Conjunction	c	Conjunction
	cc	juxtaposition conjunction
Preposition	p	Preposition
	pba	preposition ba
	pbei	preposition bei
Time Expression	t	time expression
	tg	monosyllabic morpheme as time expression
Location Word	s	location word
Direction Word	f	direction word
Non-defined letters or characters	x	non-defined letters or characters
Onomatopoeia	o	Onomatopoeia
Pronoun	r	Pronoun
	rr	personal pronoun
	rz	indicative pronoun
	rzt	time pronoun
	rzs	location pronoun
	rzv	predicate pronoun
	ry	question pronoun
	ryt	time question pronoun
	rys	location question pronoun
	ryv	predicate question pronoun
	rg	monosyllabic morpheme as pronoun
Auxiliary	u	Auxiliary

	uzhe	Aspect
	ule	perfect; past tense
	uGuo	present past tense
	ude1	of /'s
	ude2	to turn adjectives into adverbs
	ude3	degree modifier
	usuo	indicating possessive relationships
	udeng	etc.
	uyy	Like
	udh	Postponed
	uls	in terms of
	uzhi	of/'s
	ulian	Even
Number	m	Number
	mq	number classifier
	mg	monosyllabic morpheme as number
Classifier	q	Classifier
	qv	verbal classifier
	qt	temporal classifier
Prefix	h	Prefix
Suffix	k	Suffix
Modal Particle	y	modal particle
State or condition word	z	State or condition word
Interjection	e	Interjection

The number of classes is too large for our purposes, hence we reduce their number in the following categories:

Adj = Adjective
Av = Adverb
B = Differentiating word, e.g. *women doctor* is "women/b doctor/n"; there is no concept of compound in Chinese, "women/b doctor/n" will be regarded as two independent words. In most Chinese studies, differentiating word will be regarded as adjective.
C = Conjunction
K = Suffix.
N = Noun
Num = Numeral
P = Preposition [rather Pp]
Q = Classifier. Classifier is common in Chinese, some African and Austronesian languages, etc. and represents a measurement unit of the noun following it.

R = Pronoun (rather Pn)
S = Location word, like "在/p 当地/s" (In local areas)
T = Time expression
U = Auxiliaries. include: uzhe(aspect), ule(perfect; past tense), uGuo(present past tense), ude1(of /'s), ude2(to turn adjectives into adverbs), ude3 (degree modifier),usuo (indicating possessive relationships), udeng(etc.), uyy(like), udh(Dehua/postponed), uls(in terms of), uzhi(Zhi/of), ulian (Lian/ even)
V = Verb
Y = modal particle
c = Interjection
f = Direction word
h = Prefix
o = Onomatopoeia
x = non-defined letters or characters
z = condition or state of affairs, like "短短/z 几/m 年/q (short years)"

The strings of five Chinese journalistic texts are presented in the Appendix. The rank-frequencies of POS and the fittings by the right-truncated zeta distribution are displayed in Tables 2.4.5 and 2.4.6

Table 2.4.5
Rank-frequencies of POS in the Chinese Text 1

Rank	1	2	3	4	5	6	7	8	9	10
1	12	20	12	14	13	11	18	12	10	16
2	6	5	7	11	12	9	10	10	8	9
3	6	4	6	4	3	6	3	3	7	3
4	6	4	5	3	2	5	2	2	2	2
5	4	3	5	3	2	2	2	2	2	1
6	3	2	3	2	2	2	1	2	2	1
7	2	2	2	1	2	2	1	2	2	1
8	2	1	1	1	1	2	1	1	1	1
9	1	1	1	1	1	2	1	1	1	1
10				1	1		1	1	1	1
11				1	1			1	1	1
12					1			1	1	
13					1			1		
14								1		
Sum	42	42	42	42	42	41	40	40	38	37
WCS	.2857	.4762	.2857	.3333	.3095	.2683	.45	.30	.2632	.4324

Rank	11	12	13	14	15	16	17	18	19	20
1	13	12	13	10	9	7	11	10	9	6
2	8	9	7	6	5	5	6	4	5	4
3	4	7	4	6	3	3	3	3	5	3
4	3	2	3	2	3	3	2	2	2	3
5	2	1	2	2	2	2	2	1	1	3
6	2	1	1	2	1	2	1	1	1	2
7	1	1	1	1	1	1		1	1	1
8	1	1	1	1	1	1		1		1
9	1	1	1	1	1	1		1		1
10	1		1		1			1		
11	1				1					
Sum	37	35	34	31	28	25	25	25	24	24
WCS	.3514	.3429	.3824	.3226	.3214	.28	.44	.40	.375	.25

Table 2.4.6
Fitting the zeta distribution to rank-frequencies in columns of the Chinese Text 1

	Column						
R	1	2	3	4	5	6	7
1	12.18	18.37	12.53	16.34	15.38	12.68	19.01
2	7.05	7.63	7.12	7.22	6.91	7.02	7.11
3	5.13	4.56	5.11	4.48	4.32	4.96	4.00
4	4.09	3.16	4.04	3.19	3.10	3.88	2.66
5	3.43	2.38	3.37	2.45	2.40	3.21	1.94
6	2.97	1.89	2.90	1.98	1.94	2.75	1.50
7	2.63	1.56	2.56	1.65	1.63	2.41	1.20
8	2.37	1.31	2.29	1.41	1.40	2.15	1.00
9	2.16	1.13	2.08	1.22	1.22	1.94	0.84
10				1.09	1.07		0.73
11				0.97	0.96		
12					0.87		
13					0.80		
a	0.7875	1.2688	0.8165	1.1786	1.1548	0.8540	1.4177
R	9	9	9	11	13	9	10
DF	6	6	6	7	8	6	5
X^2	2.13	1.72	2.62	2.92	5.29	2.06	1.94
P	0.91	0.94	0.86	0.89	0.73	0.91	0.86

	Column							
R	8	9	10	11	12	13	14	15
1	13.49	12.22	16.54	14,07	14,14	13,72	11,00	9,24
2	6.34	6.12	6.47	6,33	6,32	5,99	5,48	4,63
3	4.08	4.08	3.74	3,97	3,94	3,69	3,65	3,09
4	2.98	3.06	2.53	2,85	2,82	2,62	2,73	2,32
5	2.34	2.45	1.87	2,21	2,18	2,01	2,18	1,86
6	1.92	2.04	1.46	1,79	1,76	1,61	1,82	1,55
7	1.62	1.75	1.19	1,50	1,47	1,34	1,56	1,33
8	1.40	1.53	0.99	1,28	1,26	1,14	1,36	1,16
9	1.23	1.36	0.84	1,12	1,10	0,99	1,21	1,03
10	1.10	1.23	0.73	0,99		0,88		0,93
11	0.99	1.12	0.64	0,90				0,85
12	0.90	1.02						
13	0.83							
14	0.76							
a	1.0883	0.9980	1.3541	1,1513	1,1621	1,1945	1,0041	0,9964
R	14	12	11	11	9	10	9	11
DF	9	9	6	7	6	6	6	7
X^2	3.30	3.89	2.14	0,82	5,25	0,64	2,22	0,58
P	0.95	0.92	0.91	0,997	0,51	0,996	0,90	0,999

	Column				
R	16	17	18	19	20
1	7.66	11.47	9.76	9.53	6.33
2	4.27	5.06	4.34	4.64	3.91
3	3.03	3.14	2.70	3.04	2.95
4	2.38	2.23	1.93	2.26	2.42
5	1.97	1.72	1.48	1.79	2.07
6	1.69	1.38	1.20	1.48	1.82
7	1.48		1.00	1.26	1.64
8	1.33		0.86		1.49
9	1.20		0.74		1.38
a	0.8434	1.1805	1.1711	1.0388	0.6946
R	9	6	9	7	9
DF	6	3	4	4	6
X^2	0.67	0.38	0.34	1.90	1.11
P	0.995	0.94	0.987	0.75	0.98

Table 2.4.7
Rank-frequencies of POS in columns of the Chinese Text 2

Rank	1	2	3	4	5	6	7	8	9	10
1	16	28	28	25	25	21	18	21	20	18
2	15	15	14	15	22	16	16	16	14	12
3	12	6	7	8	5	6	7	6	6	6
4	10	5	6	6	5	6	6	6	5	6
5	9	3	5	3	4	4	5	3	4	5
6	4	3	3	3	1	3	3	2	3	3
7	2	3	2	3	1	3	3	2	3	3
8	2	3	2	2	1	2	1	2	2	2
9	1	2	1	2	1	2	1	1	2	2
10	1	2	1	2	1	2	1	1	2	1
11	1	2	1	1	1	1	1	1	1	1
12	1	1	1	1	1	1	1	1	1	1
13	1	1	1	1	1	1	1	1	1	1
14	1	1	1	1	1	1	1	1	1	1
15	1	1	1	1	1	1	1	1	1	1
16		1			1		1	1		1
17							1	1		
18							1	1		
Sum	77	77	74	74	72	70	69	68	66	64
WCS	.2078	.3636	.3784	.3378	.3472	.30	.2609	.3088	.3030	.2812

Rank	11	12	13	14	15	16	17	18	19	20
1	19	17	15	13	12	16	12	13	11	11
2	11	12	15	12	9	6	8	9	5	6
3	6	10	6	6	7	3	8	4	3	3
4	5	4	3	6	3	3	1	3	2	3
5	3	2	2	4	3	2	1	2	2	3
6	2	1	2	2	2	1	1	2	2	2
7	2	1	1	1	1	1	1	1	2	1
8	2	1	1	1	1	1	1	1	2	1
9	2	1	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1		1	1
11	1	1	1	1	1	1	1		1	1
12	1	1	1		1	1	1		1	1
13	1	1	1			1			1	
14	1	1	1						1	
15	1	1	1							
16	1									
17	1									
Sum	60	55	52	4	42	38	37	36	35	34
WCS	.3167	.3091	.2885	.2708	.2857	.4211	.3243	.3611	.3143	.3235

Table 2.4.8
Fitting the zeta distribution to rank-frequencies in columns of the Chinese Text 2

	Column									
R	1	2	3	4	5	6	7	8	9	10
1	22.51	28.63	29.59	27.51	30.47	24.51	22.89	24.61	22.50	16.54
2	11.49	12.30	12.15	11.94	11.85	11.10	10.46	10.61	10.38	6.47
3	7.73	7.51	7.22	7.32	6.82	6.98	6.61	6.49	6.60	3.74
4	5.84	5.29	4.99	5.18	4.61	5.03	4.78	4.57	4.79	2.53
5	4.70	4.03	3.74	3.97	3.40	3.90	3.71	3.49	3.73	1.87
6	3.94	3.23	2.96	3.18	2.65	3.16	3.02	2.80	3.05	1.46
7	3.39	2.67	2.43	2.64	2.15	2.65	2.54	2.32	2.56	1.19
8	2.98	2.27	2.05	2.25	1.79	2.28	2.18	1.97	2.21	0.99
9	2.65	1.99	1.76	1.95	1.53	1.99	1.91	1.71	1.94	0.84
10	2.40	1.73	1.54	1.72	1.32	1.76	1.70	1.50	1.72	0.73
11	2.18	1.54	1.36	1.53	1.16	1.58	1.52	1.34	1.55	0.64
12	2.01	1.39	1.22	1.38	1.03	1.43	1.38	1.21	1.41	
13	1.86	1.26	1.10	1.25	0.93	1.31	1.26	1.09	1.29	
14	1.73	1.15	1.00	1.15	0.84	1.20	1.16	1.00	1.18	
15	1.61	1.06	0.91	1.05	0.76	1.11	1.07	0.92	1.10	
16		0.98			0.70		1.00	0.85		
17							0.93	0.79		
18							0.87	0.74		
a	0.973	1.218	1.284	1.204	1.362	1.142	1.130	1.213	1.116	1.354
R	0	4	4	5	4	6	0	7	0	1
D	15	16	15	15	16	15	18	18	15	11
F	12	12	11	12	11	12	13	12	12	6
X^2	17.03	1.83	1.74	1.93	12.81	3.58	6.59	4.85	2.18	2.14
P	0.15	0.999	0.999	0.999	0.31	0.99	0.92	0.96	0.999	0.91

	Column									
R.	11	12	13	14	15	16	17	18	19	20
1	20.28	20.35	19.41	16.50	14.48	14.94	13.59	14.60	10.14	11.17
2	9.23	8.86	8.40	8.03	6.90	6.35	6.18	6.50	5.26	5.51
3	5.82	5.45	5.14	5.27	4.47	3.85	3.90	4.05	3.59	3.65
4	4.20	3.86	3.63	3.91	3.29	2.70	2.81	2.89	2.73	2.72
5	3.26	2.95	2.77	3.10	2.59	2.05	2.18	2.23	2.21	2.17
6	2.65	2.37	2.23	2.56	2.13	1.64	1.77	1.80	1.86	1.80
7	2.22	1.97	1.85	2.18	1.81	1.35	1.49	1.51	1.61	1.54
8	1.91	1.68	1.57	1.90	1.57	1.15	1.28	1.29	1.42	1.34
9	1.67	1.46	1.36	1.68	1.38	0.99	1.12	1.12	1.27	1.19
10	1.48	1.28	1.20	1.51	1.24	0.87	0.99		1.15	1.07
11	1.33	1.15	1.07	1.37	1.12	0.78	0.89		1.05	0.97

12	1.21	1.03	0.96		1.02	0.70	0.81		0.97	0.89
13	1.10	0.94	0.87			0.63			0.90	
14	1.01	0.86	0.80						0.84	
15	0.94	0.79	0.74							
16	0.87									
17	0.81									
a	1.135	1.199	1.208	1.039	1.068	1.233	1.137	1.167	0.945	1.019
R	9	7	8	3	7	7	0	1	5	3
D	17	15	15	11	12	13	12	9	14	12
F	12	10	10	8	9	7	7	6	9	8
X^2	1.20	7.63	7.57	5.93	3.32	1.08	7.46	1.43	0.86	0.85
P	1.00	0.67	0.67	0.67	0.95	0.99	0.38	0.96	1.00	1.00

Table 2.4.9
Rank-frequencies of POS in columns of the Chinese Text 3

Rank	1	2	3	4	5	6	7	8	9	10
1	13	18	27	19	18	20	24	18	17	16
2	11	16	19	17	13	15	16	11	13	16
3	8	6	4	6	5	5	7	8	9	4
4	7	5	3	5	5	5	4	4	5	4
5	7	4	2	3	4	5	2	4	3	3
6	6	3	2	3	4	4	1	3	3	3
7	5	3	2	3	3	2	1	2	3	2
8	3	2	2	2	2	2	1	2	2	2
9	1	2	1	1	2		1	1	1	2
10	1	2		1	1			1	1	2
11	1	1		1	1			1		1
12		1		1	1			1		1
13					1			1		1
Sum	63	63	62	62	60	58	57	57	57	57
WCS	.2063	.2857	.4355	.3065	.30	.3448	.4211	.3158	.2982	.2807

Rank	11	12	13	14	15	16	17	18	19	20
1	21	13	14	16	14	15	16	14	14	14
2	12	12	9	15	8	12	12	11	10	7
3	6	5	5	5	4	3	2	7	5	5
4	4	4	4	5	4	3	2	3	4	3
5	4	4	4	2	3	3	2	2	3	3
6	3	3	3	2	2	2	2	1	2	3
7	2	3	3	1	2	2	2	1	1	1
8	1	3	2	1	2	1	1	1		1
9	1	3	2	1	1	1	1	1		

10	1	2	2	1	1	1	1			
11		1	2	1	1	1				
12			1		1	1				
13					1					
14					1					
Sum	55	53	51	50	45	45	41	41	39	37
WCS	.3818	.2453	.2745	.32	.3111	.3333	.3902	.3415	.3590	.3784

Table 2.4.10
Fitting the zeta distribution to rank-frequencies in columns of the Chinese Text 3

	Column						
R	1	2	3	4	5	6	7
1	15.13	21.80	30.10	23.49	20.16	21.86	27.60
2	9.32	10.36	11.23	10.43	9.64	10.71	10.33
3	7.02	6.71	6.31	6.49	6.26	7.05	5.81
4	5.74	4.93	4.19	4.63	4.60	5.24	3.86
5	4.91	3.88	3.05	3.56	3.63	4.17	2.81
6	4.33	3.19	2.35	2.88	2.99	3.45	2.17
7	3.88	2.70	1.89	2.40	2.54	2.95	1.75
8	3.54	2.34	1.56	2.06	2.20	2.57	1.45
9	3.26	2.06	1.32	1.79	1.94		1.22
10	3.03	1.84		1.58	1.74		
11	2.83	1.66		1.42	1.57		
12		1.52		1.28	1.43		
13					1.31		
a	0.6989	1.0727	1.4227	1.1714	1.0651	1.0298	1.4185
R	11	12	9	12	13	8	9
DF	8	9	6	9	10	5	6
X^2	7.06	4.36	7.50	6.06	2.90	3.17	5.20
P	0.53	0.89	0.28	0.73	0.98	0.67	0.52

	Column						
R	8	9	10	11	12	13	14
1	20.23	20.09	19.84	22.57	14.92	14.50	20.29
2	9.30	9.80	9.25	9.72	8.33	7.88	8.65
3	5.90	6.44	5.92	5.93	5.93	5.51	5.26
4	4.27	4.78	4.31	4.18	4.65	4.28	3.69
5	3.33	3.80	3.37	3.19	3.86	3.52	2.81
6	2.71	3.14	2.76	2.55	3.31	2.99	2.24
7	2.28	2.68	2.33	2.12	2.91	2.61	1.86

8	1.96	2.33	2.01	1.80	2.60	2.32	1.58
9	1.72	2.07	1.77	1.56	2.35	2.10	1.36
10	1.53	1.85	1.57	1.37	2.16	1.91	1.20
11	1.37		1.42		1.99	1.76	1.06
12	1.25		1.29			1.63	
13	1.14		1.18				
a	1.1212	1.0353	1.1011	1.2161	0.8401	0.8805	1.2290
R	13	10	13	10	11	12	11
DF	10	7	10	7	8	9	8
X^2	2.18	3.74	6.78	1.60	2.88	0.70	7.03
P	0.99	0.81	0.75	0.98	0.94	1.00	0.53

	Column					
R	15	16	17	18	19	20
1	14.58	17.22	17.55	16.89	15.54	14.10
2	7.05	7.58	7.27	7.42	7.54	6.84
3	4.60	4.69	4.34	4.58	4.94	4.48
4	3.40	3.34	3.01	3.26	3.66	3.32
5	2.69	2.57	2.27	2.50	2.90	2.63
6	2.22	2.07	1.80	2.01	2.39	2.17
7	1.89	1.72	1.48	1.68	2.04	1.85
8	1.64	1.47	1.25	1.43		1.61
9	1.45	1.28	1.08	1.24		
10	1.30	1.13	0.94			
11	1.18	1.01				
12	1.07	0.91				
13	0.99					
14	0.91					
a	1.0496	1.1831	1.2711	1.1873	1.0437	1.0438
R	14	12	10	9	7	8
DF	10	8	6	6	4	5
X^2	0.72	3.85	5.10	4.58	1.59	1.08
P	1.00	0.87	0.53	0.60	0.81	0.96

Table 2.4.11
Rank-frequencies of POS in Chinese Text 4

	Column									
Rank	1	2	3	4	5	6	7	8	9	10
1	22	25	17	21	22	25	18	17	17	14
2	9	12	13	21	13	10	12	13	11	11
3	8	4	6	4	5	7	5	5	3	7
4	3	3	3	3	4	4	5	4	2	3

5	3	3	3	2	4	2	3	2	2	2
6	3	2	3	1	2	2	3	2	2	2
7	2	2	2	1	1	1	1	2	2	2
8	2	2	2	1	1	1	1	1	2	1
9	2	1	2	1	1	1	1	1	1	1
10	1	1	2	1	1	1	1	1	1	1
11	1	1	2	1	1	1	1	1	1	1
12	1	1	1	1	1	1	1	1	1	1
13	1	1	1	1	1	1	1	1	1	1
14	1	1	1	1	1		1	1	1	1
15	1	1	1				1	1	1	
16	1	1	1						1	
17			1						1	
Sum	61	61	61	61	58	57	55	53	50	48
WCS	.3607	.4098	.2787	.3443	.3783	.4386	.3273	.3208	.34	.2917

Rank	11	12	13	14	15	16	17	18	19	20
1	16	17	18	10	9	9	12	9	10	8
2	15	7	12	8	8	7	7	9	5	8
3	2	5	3	5	3	5	2	2	2	2
4	2	3	2	5	3	3	2	2	2	1
5	2	2	1	3	3	2	2	1	2	1
6	2	2	1	1	3	2	2	1	2	1
7	1	2	1	1	2	2	1	1	1	1
8	1	1	1	1	1	1	1	1	1	1
9	1	1	1	1	1	1	1	1	1	
10	1	1	1	1	1	1	1	1		
11	1	1		1	1	1	1			
12	1	1		1	1	1	1			
13	1	1		1	1	1				
14	1	1			1					
Sum	47	45	41	39	38	36	33	28	26	23
WCS	.3404	.3778	.4390	.2564	.2368	.25	.3636	.3214	.3846	.3478

Table 2.4.12
Fitting the zeta distribution to rank-frequencies in columns of the Chinese Text 4

	Column						
R	1	2	3	4	5	6	7
1	21.95	24.57	19.14	26.50	23.74	25.41	20.06
2	9.66	9.94	9.14	10.11	9.66	9.71	8.83
3	5.97	5.86	5.93	5.75	5.71	5.53	5.46
4	4.25	4.02	4.36	3.85	3.93	3.71	3.88
5	3.26	3.01	3.44	2.83	2.94	2.72	2.98

6	2.63	2.37	2.83	2.19	2.32	2.11	2.40
7	2.19	1.94	2.40	1.77	1.90	1.71	2.00
8	1.87	1.63	2.08	1.47	1.60	1.42	1.71
9	1.63	1.40	1.84	1.25	1.37	1.20	1.49
10	1.43	1.22	1.64	1.08	1.20	1.04	1.31
11	1.28	1.07	1.48	0.94	1.06	0.91	1.17
12	1.16	0.96	1.35	0.84	0.94	0.81	1.06
13	1.05	0.86	1.24	0.75	0.85	0.72	0.96
14	0.96	0.78	1.15	0.68	0.77		0.88
15	0.89	0.72	1.06				0.81
16	0.82	0.66	0.99				
17			0.93				
a	1.1847	1.3054	1.0667	1.3906	1.2974	1.3880	1.1845
R	16	16	17	14	14	13	15
DF	11	10	13	9	9	8	10
X^2	1.55	1.94	2.87	15.30	2.67	1.22	2.97
P	1.00	1.00	1.00	0.08	0.98	1.00	0.98

	Column						
R	8	9	10	11	12	13	14
1	19.61	17.15	16.88	18.73	16.59	19.88	12.14
2	8.54	7.73	7.71	7.79	7.32	7.29	6.11
3	5.25	4.85	4.87	4.66	4.54	4.05	4.09
4	3.72	3.48	3.52	3.24	3.23	2.67	3.08
5	2.84	2.69	2.74	2.44	2.48	1.93	2.47
6	2.29	2.18	2.23	1.94	2.00	1.48	2.06
7	1.90	1.83	1.87	1.59	1.67	1.19	1.77
8	1.62	1.57	1.61	1.35	1.43	0.98	1.55
9	1.41	1.37	1.41	1.16	1.24	0.83	1.38
10	1.24	1.21	1.25	1.01	1.10	0.71	1.24
11	1.10	1.09	1.12	0.90	0.98		1.13
12	0.99	0.98	1.02	0.81	0.88		1.04
13	0.90	0.90	0.93	0.73	0.80		0.96
14	0.83	0.82	0.85	0.66	0.74		
15	0.76	0.76					
16		0.71					
17		0.66					
a	1.1998	1.1501	1.1308	1.2663	1.1801	1.4478	0.9899
R	15	17	14	14	14	10	13
DF	10	11	10	9	9	5	9
X^2	3.52	3.60	3.57	9.80	0.58	4.43	3.72
P	0.97	0.98	0.96	0.37	1.00	0.49	0.93

	Column					
R	15	16	17	18	19	20
1	10.51	10.55	12.02	10.95	9.69	9.70
2	5.61	5.52	5.50	4.91	4.64	4.29
3	3.89	3.78	3.48	3.08	3.02	2.66
4	3.00	2.89	2.52	2.21	2.22	1.90
5	2.45	2.35	1.96	1.70	1.75	1.461
6	2.08	1.98	1.59	1.38	1.44	1.18
7	1.81	1.71	1.34	1.16	1.23	0.98
8	1.60	1.51	1.15	0.99	1.06	0.84
9	1.44	1.36	1.01	0.86	0.94	
10	1.31	1.23	0.90	0.76		
11	1.20	1.12	0.80			
12	1.11	1.04	0.73			
13	1.03	0.96				
14	0.96					
a	0.9048	0.9336	1.1275	1.1558	1.0625	1.1775
R	14	13	12	10	9	8
DF	10	9	7	5	5	4
X^2	2.47	1.44	1.50	4.63	0.69	4.29
P	0.99	1.00	0.98	0.46	0.98	0.37

Table 2.4.13
Rank-frequencies of POS in Chinese Text 5

	Column									
Rank	1	2	3	4	5	6	7	8	9	10
1	15	18	18	22	24	16	16	14	17	15
2	8	16	14	14	8	12	11	12	9	10
3	7	4	4	4	4	5	5	6	4	3
4	7	4	4	3	4	4	4	5	3	3
5	5	3	3	3	3	2	3	2	3	3
6	3	3	3	1	2	2	2	2	3	2
7	3	2	1	1	2	2	1	2	3	2
8	2	1	1	1	2	1	1	2	2	2
9	1	1	1	1	1	1	1	1	1	2
10	1		1	1	1	1	1	1	1	2
11			1			1	1	1	1	1
12						1	1			1
13						1	1			
14							1			
Sum	53	54	54	55	56	55	55	56	56	56
WCS	.2830	.3333	.3333	.4	.4286	.2909	.2909	.25	.3036	.2679

	Column									
Rank	11	12	13	14	15	16	17	18	19	20
1	14	13	12	11	11	16	12	9	14	8
2	14	10	11	10	8	5	8	8	7	8
3	5	8	5	5	4	5	5	7	2	2
4	4	4	4	5	4	3	3	4	2	2
5	3	3	3	3	4	2	2	2	2	2
6	2	2	3	2	2	1	2	1	1	2
7	1	2	2	2	2	1	1	1	1	1
8	1	1	1	2	1	1	1			1
9	1	1	1	1	1	1				1
10	1	1	1	1	1	1				
11			1	1	1	1				
12					1					
Sum	46	45	44	43	40	37	34	32	29	27
WCS	.3043	.2889	.2727	.2558	.275	.4324	.3529	.2812	.4829	.2969

Table 2.4.14
Fitting the zeta distribution to rank-frequencies in columns of the Chinese Text 5

	Column						
R	1	2	3	4	5	6	7
1	15.67	21.45	20.91	24.36	22.76	18.58	18.05
2	8.57	9.41	8.84	9.07	9.07	8.13	7.97
3	6.03	5.81	5.34	5.09	5.29	5.01	4.94
4	4.69	4.12	3.74	3.37	3.61	3.55	3.52
5	3.86	3.16	2.83	2.45	2.69	2.72	2.71
6	3.30	2.55	2.26	1.89	2.11	2.19	2.18
7	2.88	2.12	1.87	1.52	1.72	1.82	1.82
8	2.57	1.81	1.58	1.26	1.44	1.55	1.55
9	2.32	1.57	1.39	1.06	1.23	1.35	1.35
10	2.11		1.20	0.91	1.07	1.19	1.19
11			1.06			1.06	1.07
12						0.96	0.96
13						0.87	0.88
14							0.80
a	0.8700	1.1895	1.2418	1.4264	1.3274	1.1932	1.1791
R	10	9	11	10	10	13	14
DF	7	6	8	6	7	9	9
X^2	3.19	6.41	4.77	3.95	0.91	2.83	2.25
P	0.87	0.38	0.78	0.68	1.00	0.97	0.99

	Column						
R	8	9	10	11	12	13	14
1	17.15	16.92	15.34	18.16	15.74	14.85	13.63
2	8.11	7.95	7.49	8.08	7.73	7.32	7.03
3	5.23	5.11	4.92	5.09	5.10	4.84	4.77
4	3.83	3.74	3.66	3.60	3.79	3.61	3.62
5	3.01	2.93	2.90	2.77	3.02	2.88	2.93
6	2.47	2.40	2.40	2.24	2.50	2.39	2.46
7	2.09	2.03	2.05	1.87	2.14	2.04	2.12
8	1.81	1.76	1.78	1.60	1.86	1.78	1.87
9	1.59	1.54	1.58	1.40	1.65	1.58	1.67
10	1.42	1.38	1.42	1.23	1.48	1.42	1.51
11	1.28	1.24	1.28			1.29	1.38
12			1.17				
a	1.0814	1.0896	1.0347	1.1675	1.0267	1.0199	0.9556
R	11	11	12	10	10	11	11
DF	8	8	9	7	7	8	8
X^2	3.78	1.51	2.26	6.16	3.73	3.35	2.95
P	0.88	0.99	0.99	0.52	0.81	0.91	0.94

	Column					
R	15	16	17	18	19	20
1	12.36	15.37	13.39	10.88	14.34	9.59
2	6.35	6.43	6.31	6.07	5.54	4.78
3	4.31	3.86	4.06	4.32	3.18	3.18
4	3.27	2.69	2.97	3.39	2.14	2.38
5	2.64	2.03	2.33	2.81	1.58	1.90
6	2.22	1.61	1.91	2.41	1.23	1.58
7	1.91	1.33	1.62	2.12	0.99	1.36
8	1.68	1.12	1.40			1.19
9	1.50	0.97				1.05
10	1.36	0.85				
11	1.24	0.75				
12	1.14					
a	0.9591	1.2587	1.0858	0.8411	1.3719	1.0054
R	12	11	8	7	7	9
DF	9	6	5	4	3	6
X^2	2.09	1.14	1.22	4.36	0.97	3.18
P	1.00	0.98	0.94	0.36	0.81	0.79

For the individual Chinese texts we obtain the graphs of the weighted consensus string as shown in Fig. 2.4.2.

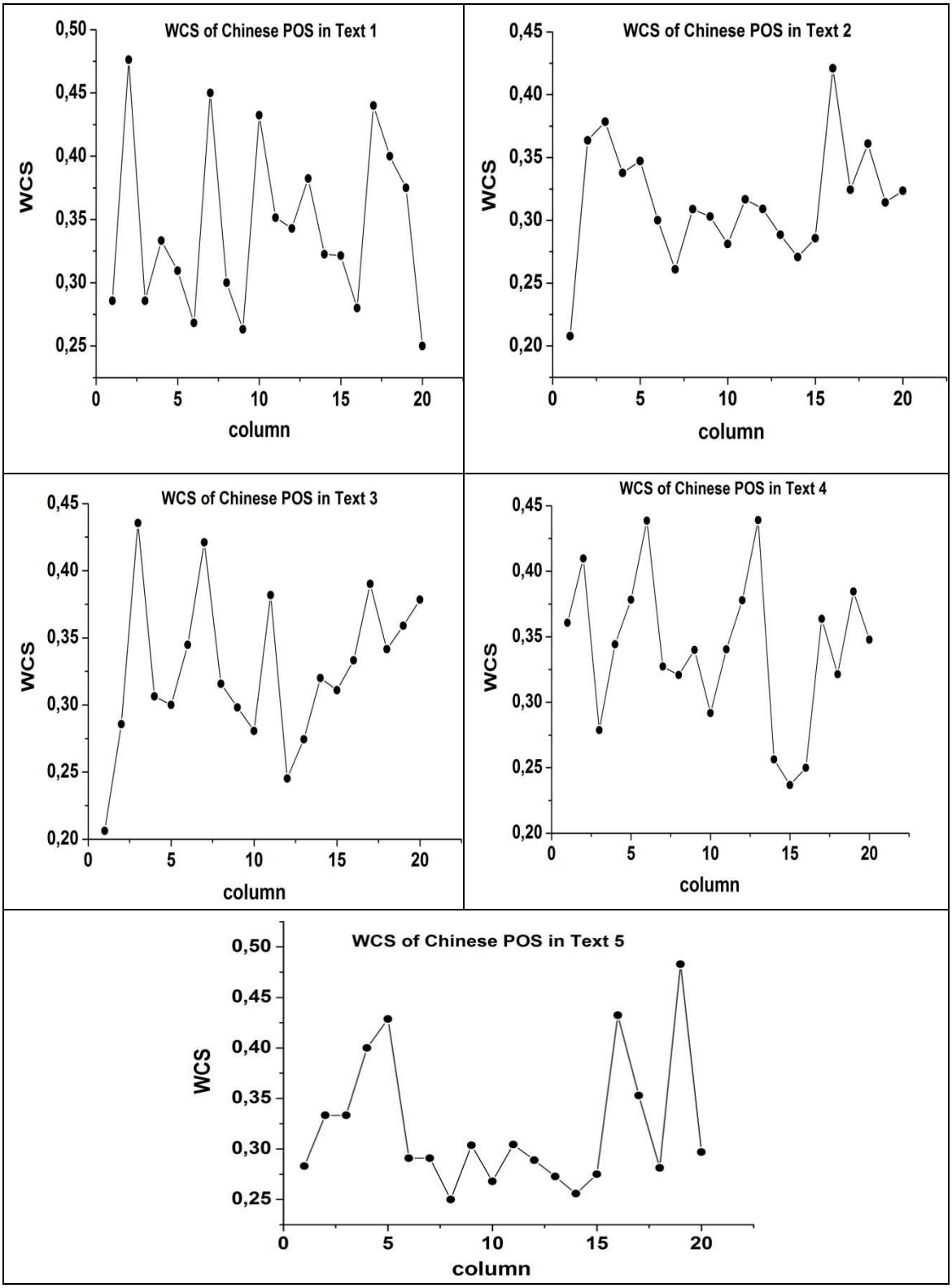


Figure 2.4.2. WCS of POS in Chinese texts

As for **Polish,** grammars tend to distinguish the parts of speech using a mixture of semantic, morphological, and functional criteria. This is often a convenient practice for discussing Polish by itself, but it may yield avoidable incompatibilities when other languages are involved. In this work, we focus on the functional criterion, hence only the following classes will be distinguished:

- ADJ[ective], including some numerals (e.g. *lat dziewięćdziesiątych* 'of the 90s'), some participles (e.g. *uaktywniający* 'activating'), and some pronouns (e.g. *takie* 'such a … *neuter sg.*');
- ADV[erb], including some numerals (e.g. *kilkakrotnie* 'several times'), some participles (e.g. *kupując* '(when) buying'), and some pronouns (e.g. *bardzo* '(very) much');
- CONJ[unction] (e.g. *i* 'and');
- N[oun], including deverbal nouns (e.g. *pytanie* 'question(ing)'), some numerals (e.g. *jedni* 'some (people)', lit. 'ones'), proper nouns, some pronouns (e.g. *to* 'this *neuter sg.*'), and verbs in the infinitive,
- PART[icle], including negation (e.g. *nie* 'not, non', *przecie* 'after all');
- POST[position] (e.g. *te* 'too, also');
- PREP[osition] (e.g. *w* 'in');
- V[erb] (conjugated); and
- V[erbal]PRON[oun] (*się* with verbs).[2]

Here we present only the rank-frequencies of **Polish** POS in individual positions from 1 to 15. The results are presented in Tables 2.4.15 to 2.4.24. As usual, we fit the ranked data by means of the right truncated zeta distribution.

Table 2.4.15
Rank-frequencies of POS in Polish Text 1

	Column							
Rank	1	2	3	4	5	6	7	8
1	16	21	18	22	18	16	22	15
2	11	10	9	7	10	8	7	11
3	9	7	9	5	7	8	6	9
4	7	6	4	5	7	8	6	4
5	4	2	4	4	4	4	3	3
6	3	2	3	4	2	3	2	2
7		1	2	2	1	1	1	1
8		1	1	1		1		

2 The primary function of the word *się* is to turn a transitive verb into a reflexive one, e.g. *myć* 'to wash' : *myć się* 'to wash oneself', but it is also used in semantically less clear formations, e.g. 'to laugh' can only be translated into *śmiać się*, never *śmiać* alone. In this work, *się* has been given its own, separate part of speech in order to facilitate comparison with other languages; the same word is present e.g. in Russian (*-ся*) or French (*se*), but since it follows rather different distribution patterns there, it will be hopefully easier to capture this way.

Sum	50	50	50	50	49	49	47	45
WCS	.32	.42	.36	.44	.37	.33	.47	.33

	Column						
Rank	9	10	11	12	13	14	15
1	23	18	22	10	19	13	9
2	7	8	6	8	6	8	6
3	5	5	3	6	4	3	5
4	4	3	3	5	3	3	4
5	4	3	2	4	2	2	3
6	1	2	1	4	1	2	1
7		1	1		1	1	1
8							1
Sum	44	40	38	37	36	32	30
WCS	.52	.45	.58	.27	.52	.41	.30

Table 2.4.16
Fitting the RT-Zeta distribution to the Polish POS data in Text 1

	Column							
R	1	2	3	4	5	6	7	8
1	17.06	21.80	18.73	20.30	18.54	16.30	21.00	17.15
2	10.10	9.33	9.22	9.30	9.43	8.85	9.09	8.67
3	7.43	5.68	6.09	5.89	6.35	6.19	5.57	5.82
4	5.97	3.99	4.54	4.26	4.80	4.80	3.93	4.38
5	5.05	3.04	3.62	3.32	3.86	3.95	3.00	3.52
6	4.40	2.43	3.00	2.70	3.23	3.36	2.41	2.94
7		2.01	2.56	2.27	2.78	2.93	2.00	2.53
8		1.71	2.24	1.95		2.61		
a	0.7568	1.2243	1.0220	1.1255	0.9752	0.8812	1.2081	0.9844
R	6	8	8	8	7	8	7	7
DF	3	5	5	5	4	5	4	4
X^2	1.32	2.63	2.33	2.24	2.74	5.05	2.22	3.97
P	0.73	0.76	0.80	0.82	0.60	0.41	0.70	0.41

	Column						
R	9	10	11	12	13	14	15
1	21.66	18.19	21.24	10.53	18.66	13.76	9.95
2	8.80	7.73	7.01	7.24	6.81	6.20	5.41
3	5.19	4.68	3.67	5.82	3.77	3.88	3.79
4	3.57	3.28	2.31	4.98	2.48	2.79	2.95
5	2.67	2.49	1.62	4.42	1.79	2.16	2.42
6	2.11	1.99	1.21	4.01	1.38	1.75	2.07
7		1.64	0.95		1.10	1.46	1.80
8							1.60
a	1.3006	1.2357	1.5992	0.5392	1.4551	1.1513	0.8774
R	6	7	7	6	7	7	8
DF	3	4	3	3	4	4	5
X^2	1.75	0.41	0.60	0.15	0.36	0.98	2.19
P	0.63	0.98	0.90	0.99	0.99	0.91	0.82

Table 2.4.17
Rank-frequencies of POS in Polish Text 2

	Column							
Rank	1	2	3	4	5	6	7	8
1	12	26	20	29	20	15	18	15
2	11	7	9	8	12	14	10	8
3	11	6	7	4	7	8	7	7
4	8	5	5	3	4	7	5	6
5	6	4	4	3	3	3	5	5
6	2	2	4	2	2	2	2	4
7	1	1	1	1	1		1	1
Sum	51	51	50	50	49	49	48	46
WCS	.24	.51	.40	.58	.41	.31	.38	.33

	Column						
Rank	9	10	11	12	13	14	15
1	19	23	14	19	12	12	15
2	9	5	13	11	9	9	6
3	8	5	5	4	5	5	4
4	5	3	4	3	4	4	1
5	2	3	4	1	4	2	1
6	1	3	2	1	3	1	1
7	1	1		1			1
8		1					
Sum	45	44	42	40	37	33	29
WCS	.42	.52	.33	.48	.32	.36	.52

Table 2.4.18
Fitting the RT-Zeta distribution to the Polish POS data in text 2

	Column							
R	1	2	3	4	5	6	7	8
1	14.19	24.49	19.61	27.83	21.87	17.57	18.54	14.81
2	9.27	9.79	9.65	9.24	9.47	9.95	9.26	8.64
3	7.22	5.73	6.38	4.85	5.81	7.13	6.17	6.31
4	6.05	3.92	4.75	3.07	4.10	5.63	4.62	5.05
5	5.27	2.92	3.78	2.15	3.14	4.69	3.70	4.24
6	4.71	2.29	3.14	1.61	2.52	4.04	3.08	3.68
7	4.29	1.87	2.68	1.26	2.09		2.64	3.27
a	0.6151	1.3222	1.0224	1.5909	1.2068	0.8212	1.0017	0.7765
R	7	7	7	7	7	6	7	7
DF	4	4	4	4	4	3	4	4
X^2	7.45	2.05	1.43	0.85	1.76	4.10	2.07	2.04
P	0.11	0.73	0.84	0.93	0.78	0.25	0.72	0.73

	Column						
R	9	10	11	12	13	14	15
1	19.72	20.89	16.20	20.78	12.63	13.39	15.23
2	8.71	8.18	8.56	7.56	7.47	6.74	5.46
3	5.40	4.73	5.90	4.18	5.50	4.51	3.00
4	3.85	3.21	4.53	2.75	4.42	3.39	1.96
5	2.96	2.37	3.69	1.99	3.73	2.72	1.41
6	2.38	1.85	3.12	1.52	3.25	2.27	1.08
7	1.99	1.50		1.22			0.86
8		1.26					
a	1.1791	1.3519	0.9195	1.4587	0.7574	0.9909	1.4789
R	7	8	6	7	6	6	7
DF	4	5	3	4	3	3	3
X^2	3.24	2.58	3.22	2.46	0.47	1.97	0.98
P	0.52	0.76	0.36	0.65	0.93	0.58	0.81

Table 2.4.19
Rank-frequencies of POS in Polish Text 3

	Column							
Rank	1	2	3	4	5	6	7	8
1	14	21	17	13	14	18	17	16
2	14	6	11	10	11	8	10	6
3	12	6	8	10	7	7	7	6
4	3	5	5	7	5	6	6	5

5	3	5	3	3	4	4	4	4
6	2	3	3	2	4	2	3	4
7	2	2	2	2	3	2	2	3
8		2	1	1	1	2		2
9				1				2
Sum	50	50	50	49	49	49	49	48
WCS	.28	.42	.34	.27	.29	.37	.35	.33

	Column						
Rank	9	10	11	12	13	14	15
1	11	20	14	17	10	12	15
2	11	5	8	6	10	10	8
3	9	5	6	6	8	6	7
4	6	4	5	5	6	6	3
5	4	4	5	3	4	3	2
6	4	3	4	3	1	2	1
7	1	3	1	1	1	1	1
8	1	2		1			
Sum	47	46	43	42	40	40	37
WCS	.23	.43	.33	.40	.25	.30	.41

Table 2.4.20
Fitting the right truncated zeta distribution to Polish POS in Text 3

	Column							
R	1	2	3	4	5	6	7	8
1	17.69	19.24	18.82	15.75	15.64	18.00	17.74	14.90
2	9.55	9.25	9.23	8.47	8.77	9.01	9.39	8.22
3	6.66	6.03	6.08	5.90	6.25	6.01	6.47	5.81
4	5.16	4.45	4.52	4.56	4.92	4.51	4.97	4.54
5	4.23	3.52	3.60	3.73	4.08	3.61	4.05	3.75
6	3.60	2.90	2.98	3.17	3.50	3.01	3.42	3.20
7	3.14	2.46	2.54	2.76	3.08	2.58	2.97	2.81
8		2.14	2.22	2.45	2.76	2.26		2.50
9				2.21				2.26
a	0.8882	1.0562	1.0284	0.8942	0.8350	0.9981	0.9186	0.8579
R	7	8	8	9	8	8	7	9
DF	4	5	5	6	5	5	4	6
X^2	9.51	2.10	2.06	7.23	2.02	1.31	0.90	1.10
P	0.05	0.84	0.84	0.30	0.85	0.93	0.95	0.98

	Column						
R	9	10	11	12	13	14	15
1	13.84	17.62	13.85	16.41	12.18	13.76	15.96
2	8.24	8.51	8.08	7.79	7.43	7.61	7.16
3	6.08	5.56	5.90	5.04	5.56	5.38	4.48
4	4.91	4.11	4.72	3.70	4.53	4.21	3.21
5	4.15	3.25	3.96	2.91	3.86	3.48	2.48
6	3.62	2.68	3.44	2.39	3.39	2.97	2.01
7	3.23	2.28	3.05	2.02	3.04	2.61	1.68
8	2.92	1.98		1.75			
a	0.7481	1.0500	0.7771	1.0752	0.7135	0.8549	1.1563
R	8	8	7	8	7	7	7
DF	5	5	4	5	4	4	4
X^2	5.99	2.26	1.76	2.08	5.88	3.19	2.46
P	0.31	0.81	0.78	0.84	0.21	0.53	0.65

Table 2.4.21
Rank-frequencies of POS in Polish Text 4

	Column							
Rank	1	2	3	4	5	6	7	8
1	21	16	14	13	15	14	12	15
2	8	13	11	10	9	11	12	10
3	7	5	8	7	8	9	6	6
4	4	5	6	6	8	6	5	4
5	4	3	5	5	5	3	4	3
6	3	3	3	4	2	2	4	3
7	2	2	1	2		1	2	2
8	1	1	1				1	1
9		1						
Sum	50	49	49	47	47	46	46	44
WCS	.42	.33	.29	.28	.32	.30	.26	.34

	Column						
Rank	9	10	11	12	13	14	15
1	21	14	18	17	15	12	14
2	8	13	7	8	7	8	9
3	5	4	6	5	4	5	4
4	3	3	5	4	4	4	3
5	3	3	2	3	3	4	3
6	2	3	2	2	3	3	3
7	1	2	1	1	2	2	1
8		1			1	1	1
Sum	43	43	41	40	39	39	38
WCS	.49	.33	.44	.42	.38	.31	.37

Table 2.4.22
Fitting the right truncated zeta distribution to Polish POS in Text 4

	Column							
R	1	2	3	4	5	6	7	8
1	20.65	18.68	16.08	14.08	15.04	16.19	14.55	16.41
2	9.31	8.78	8.82	8.70	9.41	8.78	8.21	8.11
3	5.85	5.64	6.21	6.56	7.15	6.14	5.88	5.37
4	4.20	4.12	4.84	5.37	5.88	4.76	4.64	4.01
5	3.25	3.23	3.99	4.60	5.06	3.91	3.86	3.20
6	2.64	2.65	3.41	4.05	4.47	3.33	3.32	2.66
7	2.21	2.24	2.98	3.64		2.90	2.92	2.27
8	1.89	1.94	2.66				2.62	1.98
9		1.70						
a	1.1487	1.0897	0.8655	0.6955	0.6771	0.8829	0.8249	1.0164
R	8	9	8	7	6	7	8	8
DF	5	6	5	4	3	4	5	5
X^2	1.09	3.51	4.25	1.16	2.25	4.51	3.66	1.21
P	0.95	0.74	0.51	0.89	0.52	0.34	0.60	0.94

	Column						
R	9	10	11	12	13	14	15
1	20.87	16.54	17.82	17.25	14.63	12.68	14.80
2	8.24	7.96	7.94	7.74	7.19	7.01	7.04
3	4.79	5.19	4.94	4.85	4.75	4.95	4.56
4	3.26	3.83	3.54	3.48	3.54	3.87	3.35
5	2.41	3.02	2.72	2.69	2.82	3.20	2.64
6	1.89	2.50	2.20	2.18	2.34	2.74	2.17
7	1.54	2.12	1.84	1.82	1.99	2.40	1.84
8		1.84			1.74	2.14	1.59
a	1.3401	1.0558	1.1667	1.1554	1.0239	0.8555	1.0717
R	7	8	7	7	8	8	8
DF	4	5	4	4	5	5	5
X^2	0.37	4.53	1.54	0.52	0.71	1.08	1.66
P	0.98	0.48	0.82	0.97	0.98	0.96	0.89

Table 2.4.23
Rank-frequencies of POS in Polish text 5

	Column							
Rank	1	2	3	4	5	6	7	8
1	13	18	15	18	21	14	19	16
2	10	13	9	12	10	10	5	8
3	9	6	9	6	5	7	5	6
4	8	4	6	4	4	6	5	5
5	4	3	6	3	4	5	4	5
6	3	3	5	2	2	4	3	4
7	3	2		2	1		3	1
8		1		2			1	
9				1				
Sum	50	50	50	50	47	46	45	45
WCS	.26	.36	.30	.36	.45	.30	.42	.36

	Column						
Rank	9	10	11	12	13	14	15
1	17	15	16	9	11	18	6
2	8	12	8	5	10	2	6
3	6	4	6	5	4	2	4
4	4	3	3	5	3	2	2
5	3	3	3	3	2	2	2
6	2	1	1	3	1	1	2
7	1	1		2	1	1	1
8				2			
Sum	41	39	37	34	32	28	23
WCS	.41	.38	.43	.26	.34	.64	.26

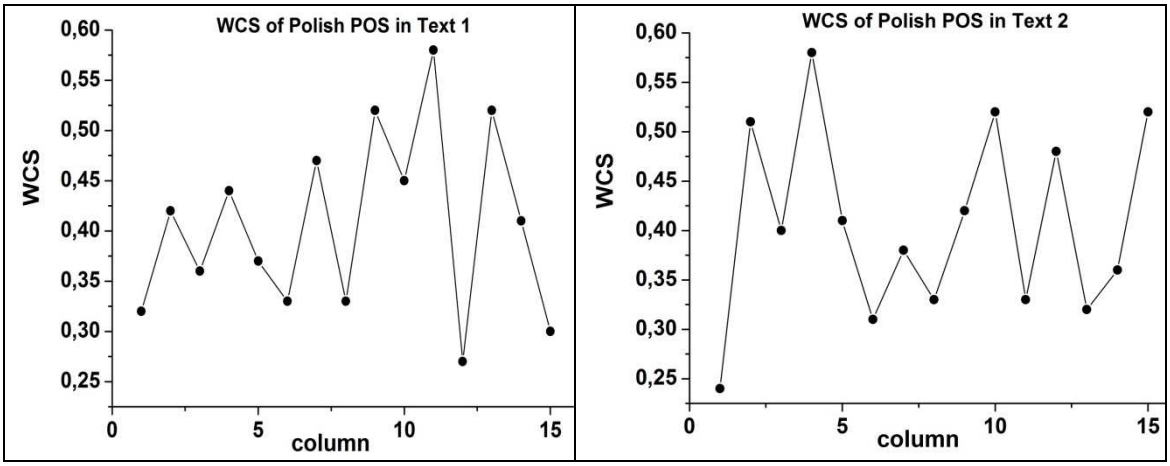
Table 2.4.24
Fitting the right truncated zeta distribution to Polish POS in Text 5

	Column							
R	1	2	3	4	5	6	7	8
1	14.41	20.23	14.80	19.98	21.31	14.46	16.95	15.49
2	9.17	9.30	9.87	9.01	9.08	9.18	8.31	8.56
3	7.03	5.90	7.78	5.66	5.51	7.04	5.47	6.05
4	5.83	4.28	6.58	4.07	3.87	5.83	4.07	4.73
5	5.04	3.33	5.77	3.15	2.94	5.03	3.23	3.91
6	4.47	2.71	5.19	2.55	2.35	4.47	2.68	3.34
7	4.05	2.28		2.14	1.94		2.29	2.93
8		1.97		1.84			1.99	
9				1.60				

a	0.6529	1.1209	0.5850	1.1480	1.2309	0.6553	1.0292	0.8562
R	7	8	6	9	7	6	8	7
DF	4	5	3	6	4	3	5	4
X^2	2.54	2.31	0.34	1.58	1.04	0.14	2.75	1.78
P	0.64	0.80	0.95	0.95	0.90	0.99	0.74	0.78

	Column						
R	9	10	11	12	13	14 Si-Po	15
1	17.39	17.23	16.47	8.97	13.28	17.46	7.27
2	7.94	7.54	7.52	5.77	6.19	1.82	4.31
3	5.02	4.65	4.75	4.46	3.97	2.61	3.17
4	3.62	3.30	3.43	3.71	2.89	2.50	2.55
5	2.81	2.53	2.67	3.22	2.26	1.79	2.15
6	2.29	2.04	2.17	2.87	1.85	1.03	1.88
7	1.92	1.70		2.60	1.56	0.80	1.67
8				2.39			
a	1.1316	1.1911	1.1312	0.6360	1.1000	2.8712	0.7559
R	7	7	6	8	7	$\alpha = 0.3991$	7
DF	4	4	3	5	4	3	4
X^2	0.73	3.94	1.10	0.84	3.36	0.32	1.51
P	0.95	0.41	0.78	0.97	0.50	0.96	0.82

In column 14 of Text 5 one was forced to apply the Singh-Poisson modification, all other columns follow the zeta distribution.



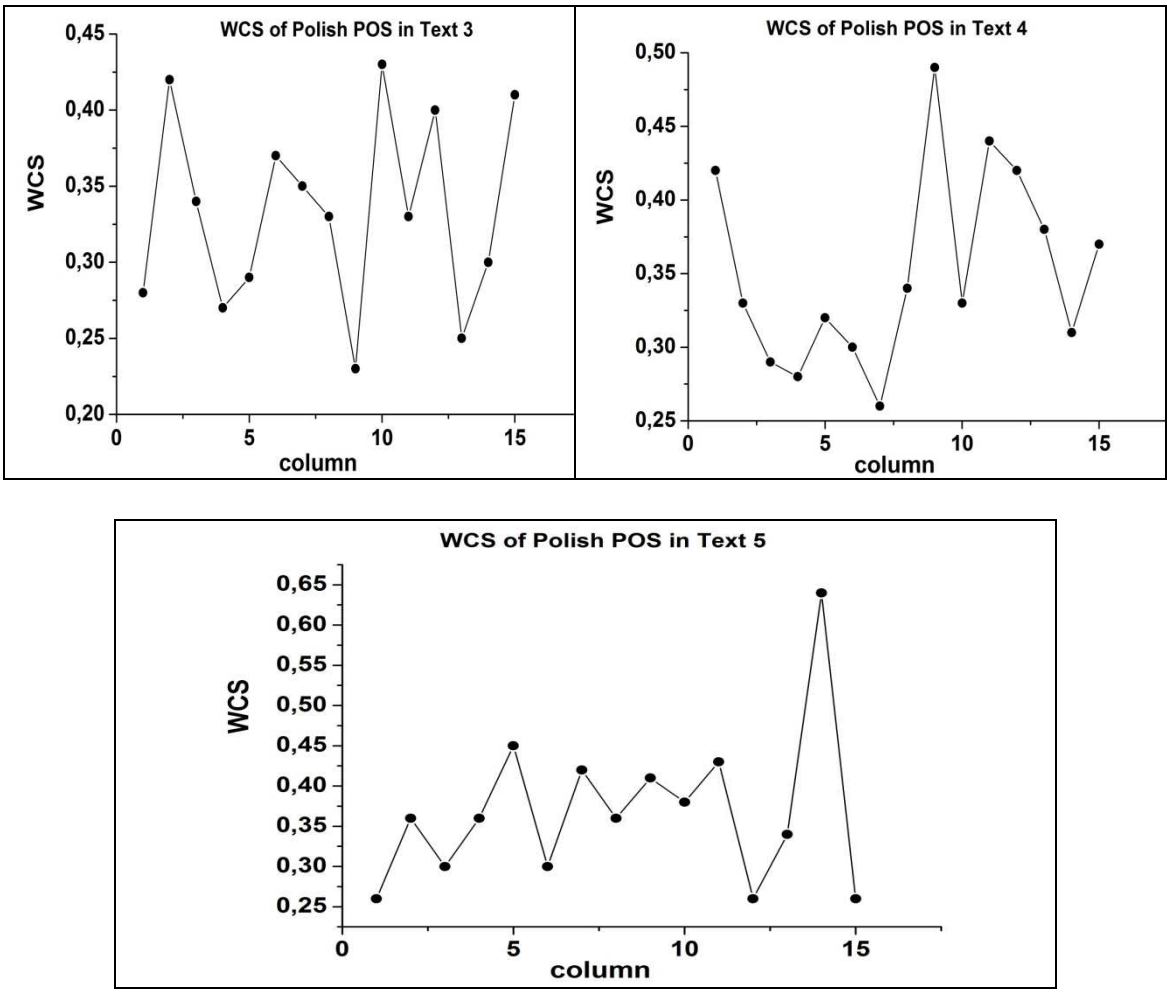


Figure 2.4.3. WCS of POS in Polish texts

For **Persian**, the following word classes have been stated:

A = adjective
ADV = adverb
P = preposition
N = noun
DET = determiner
PRO = pronoun
V = verb
CON = conjunction
AUX - auxiliary
RA = direct object identifier
IF = conditional

Though there are many other POS tags in Persian, for the purpose of this study we tried to simplify them to be more practical and concise

The individual data can be found in the Appendix. In Tables PPOSI we present the rank-frequencies of parts of speech in five Persian texts. In Tables PPOST one finds the fitting of the Hyperpoisson distribution to this data. We considered only columns with at least 20 observations. For smaller values

Table 2.4.25
Rank-frequencies of POS in Persian Text 1

	Column											
Rank	1	2	3	4	5	6	7	8	9	10	11	12
1	12	13	19	16	11	20	14	14	10	11	13	16
2	7	4	3	5	9	2	7	5	7	7	5	3
3	5	4	2	3	3	2	4	4	6	4	3	3
4	3	4	2	2	2	2	4	3	3	4	3	2
5	1	3	1	2	2	2	1	2	2	3	2	2
6	1	1	1	1	1	1		1	1	1	2	2
7	1	1	1	1	1	1		1	1		1	1
8			1		1	1					1	1
Sum	30	30	30	30	30	30	30	30	30	30	30	30
WCS	0.4	0.43	0.63	0.53	0.37	0.67	0.47	0.47	0.33	0.37	0.43	0.53

	Column								
Rank	13	14	15	16	17	18	19	20	21
1	12	9	13	10	8	13	9	6	6
2	7	7	3	4	5	3	3	5	4
3	5	4	3	3	3	2	3	3	4
4	3	3	3	3	3	1	2	2	3
5	3	2	3	2	3	1	2	2	2
6		2	1	1	1	1	2	1	1
7		1	1	1		1	1	1	1
8								1	
Sum	30	28	27	24	23	22	22	21	21
WCS	0.4	0.32	0.48	0.42	0.35	0.59	0.41	0.29	0.29

Table 2.4.26
Fitting the right truncated zeta distribution to POS in Persian Text 1

	Column					
Rank	1	2	3	4	5	7
1	12.95	12.07	16.97	15.49	12.81	14.03
2	5.81	5.80	5.36	5.68	5.60	6.49
3	3.63	3.78	2.74	3.16	3.45	4.13
4	2.60	2.79	1.90	2.08	2.45	3.00

5	2.01	2.20	1.17	1.51	1.87	2.34
6	1.63	1.82	0.87	1.16	1.51	-
7	1.36	1.54	0.67	0.93	1.25	-
8	-	-	0.54		1.07	-
a	1.1567	1.0575	1.6602	1.4473	1.1942	1.1125
R	7	7	8	7	8	5
DF	4	4	3	3	5	2
X^2	1.74	2.02	2.06	0.27	2.90	1.15
P	0.78	0.73	0.56	0.96	0.75	0.56

	Column						
Rank	8	9	10	11	12	13	14
1	13.76	11.29	11.38	12.52	14.30	12.22	10.11
2	5.79	5.77	6.11	5.59	5.58	6.62	5.36
3	3.49	3.9ß	4.25	3.49	3.22	4.63	3.70
4	2.44	2.95	3.28	2.50	2.18	3.59	2.84
5	1.84	2.38	2.69	1.93	1.61	2.94	2.32
6	1.47	1.99	2.28	1.56	1.25	-	1.96
7	1.21	1.72	-	1.30	1.02	-	1.70
8	-	-	-	1.11	0.85	-	-
a	1.2487	0.9680	0.8966	1.1631	1.3585	0.8845	0.9155
R	7	7	6	8	8	5	7
DF	4	4	3	5	4	2	4
X^2	0.52	2.49	1.07	0.46	1.97	0.15	0.99
P	0.97	0.66	0.78	0.99	0.74	0.93	0.91

	Column						
Rank	15	16	17	18	19	20	21
1	11.50	9.72	8.02	11.98	8.23	6.91	6.48
2	5.23	4.64	4.66	4.10	4.23	3.78	3.91
3	3.30	3.01	3.39	2.19	2.87	2.66	2.91
4	2.38	2.22	2.70	1.40	2.17	2.07	2.36
5	1.84	1.75	2.27	0.99	1.75	1.71	2.01
6	1.50	1.44	1.97	0.75	1.47	1.46	1.76
7	1.26	1.22	-	0.59	1.27	1.28	1.57
8	-	-	-	-	-	1.14	-
a	1.1371	1.0658	0.7840	1.5469	0.9609	0.8680	0.7275
R	7	7	6	7	7	8	7
DF	4	4	3	2	4	5	4
X^2	2.28	0.58	0.81	0.79	0.73	0.83	1.15
P	0.68	0.96	0.85	0.67	0.95	0.98	0.89

Table 2.4.27
Rank-frequencies of POS in the Persian Text 2

	Column										
Rank	1	2	3	4	5	6	7	8	9	10	11
1	9	14	16	14	9	14	12	13	10	14	13
2	8	5	5	4	6	6	6	5	6	5	4
3	6	3	4	3	5	3	4	4	5	4	4
4	3	3	3	3	4	2	3	3	3	3	4
5	2	2	1	2	2	2	2	3	2	2	1
6	1	1		2	2	1	2	1	1	1	1
7				1	1	1			1		1
8									1		1
Sum	29	29	29	29	29	29	29	29	29	29	29
WCS	0.31	0.48	0.55	0.48	0.31	0.48	0.41	0.45	0.34	0.48	0.45

		Column							
Rank	12	13	14	15	16	17	18	19	20
1	18	9	11	11	10	7	8	6	7
2	3	6	4	5	6	7	4	5	6
3	2	5	4	5	3	6	3	3	2
4	2	4	3	3	3	1	2	3	2
5	1	3	3	2	3	1	2	3	2
6	1	1	1	1	2	1	2	2	2
7	1	1	1	1		1	1		
8	1		1			1	1		
9			1						
Sum	29	29	29	28	27	25	23	22	21
WCS	0.62	0.31	0.38	0.39	0.37	0.28	0.35	0.27	0.33

Table 2.4.28
Fitting the zeta distribution to rank-frequencies of POS in Persian Text 2

	Column						
Rank	1	2	3	4	5	6	7
1	10.56	13.64	15.55	12.86	9.74	14.31	12.09
2	5.89	5.61	6.00	5.61	5.49	5.54	5.92
3	4.19	3.34	3.44	3.45	3.93	3.18	3.89
4	3.29	2.31	2.31	2.45	3.10	2.15	2.89
5	2.73	1.73	1.70	1.87	2.58	1.58	2.30
6	2.34	1.37	-	1.51	2.22	1.23	1.91
7	-	-	-	1.25	1.95	1.00	-

a	0.8411	1.2812	1.3739	1.1974	0.8263	1.3676	1.0312
R	6	6	5	7	7	7	6
DF	3	3	2	4	4	3	3
X^2	2.75	0.46	0.77	0.99	1.27	0.20	0.05
P	0.43	0.93	0.68	0.91	0.87	0.98	1.00

	Column						
Rank	8	9	10	11	12	13	14
1	12.43	10.95	13.67	12.34	16.09	9.68	10.61
2	5.91	5.35	5.85	5.41	5.22	5.49	5.16
3	3.82	3.52	3.56	3.34	2.70	3.94	3.38
4	2.81	2.62	2.50	2.37	1.69	3.11	2.51
5	2.21	2.08	1.90	1.82	1.18	2.59	1.99
6	1.82	1.72	1.52	1.46	0.88	2.23	1.65
7	-	1.47	-	1.22	0.68	1.97	1.40
8	-	1.28	-	1.04	0.55	-	1.22
9	-	-	-	-	-	-	1.08
a	1.0727	1.0318	1.2255	1.1903	1.6237	0.8189	1.0398
R	6	8	6	8	8	7	9
DF	3	5	3	5	3	4	6
X^2	0.84	1.35	0.47	2.21	1.89	1.86	1.41
P	0.84	0.93	0.93	0.82	0.60	0.76	0.97

	Column					
Rank	15	16	17	18	19	20
1	11.50	9.98	8.93	7.86	6.26	7.51
2	5.23	5.49	4.58	4.18	4.31	4.26
3	3.30	3.87	3.10	2.89	3.46	3.06
4	2.38	3.02	2.35	2.22	2.96	2.42
5	1.84	2.50	1.89	1.81	2.63	2.02
6	1.50	2.13	1.59	1.53	2.38	1.74
7	1.26	-	1.37	1.33	-	-
8	-	-	1.20	1.18	-	-
a	1.1371	0.8613	0.9639	0.9125	0.5392	0.8173
R	7	6	8	8	6	6
DF	4	3	5	5	3	3
X^2	2.28	0.35	5.97	0.31	0.30	1.22
P	0.68	0.95	0.31	1.00	0.96	0.75

Table 2.4.29
Rank-frequencies of POS in the Persian Text 3

	Column											
Rank	1	2	3	4	5	6	7	8	9	10	11	12
1	22	21	24	17	17	13	19	12	11	12	15	18
2	3	3	2	6	6	8	5	4	9	6	6	6
3	2	3	2	4	3	4	2	3	5	4	5	2
4	2	1	1	3	2	4	2	3	3	3	2	2
5	1	1	1		2	1	1	3	1	2	1	1
6	1	1					1	2		1		
7	1							2		1		
8								1				
Sum	32	30	30	30	30	30	30	30	29	29	29	29
WCS	.69	.70	.80	.57	.57	.43	.83	.40	.38	.41	.52	.62

	Column								
Rank	13	14	15	16	17	18	19	20	21
1	14	10	12	14	10	11	8	8	8
2	6	6	5	4	7	5	4	4	6
3	3	4	5	3	3	4	3	2	3
4	2	4	4	3	3	2	3	2	2
5	2	3	2	2	2	2	2	2	1
6	1	1	1	1	1	1	1	2	1
7	1	1		1		1	1	1	
8							1		
Sum	29	29	29	28	26	26	23	21	21
WCS	.48	.34	.41	.50	.38	.42	.35	.38	.38

Table 2.4.30
Fitting the zeta distribution to rank-frequencies of POS in Persian Text 3

	Column						
Rank	1	2	3 Si-Po	4	5	6	7
1	12.95	19.70	23.34	16.67	16.78	13.47	18.58
2	5.81	5.14	2.75	6.70	6.08	6.54	5.40
3	3.63	2.34	2.16	3.93	3.36	4.29	2.62
4	2.60	1.34	1.12	2.69	2.20	3.18	1.57
5	2.01	0.87	0.62		1.59	2.52	1.06
6	1.63	0.61					0.76
7	1.36						

a	1.1567	1.9388	1.5649	1.3147	1.4649	1.0419	1.7815
R	7	6	$\alpha = 0.2806$	4	5	5	6
DF	4	2	1	1	2	2	2
X^2	1.74	1.43	0.27	0.12	0.17	1.49	0.32
P	0.78	0.49	0.60	0.73	0.92	0.48	0.85

	Column						
Rank	8	9	10	11	12	13	14
1	10.78	12.30	12.46	15.03	18.04	14.31	10.33
2	5.50	6.38	5.61	6.08	5.46	5.54	5.54
3	3.71	4.34	3.52	3.58	2.72	3.18	3.85
4	2.81	3.30	2.53	2.46	1.65	2.15	2.97
5	2.26	2.67	1.96	1.84	1.13	1.58	2.43
6	1.89		1.59			1.23	2.07
7	1.63		1.33			1.00	1.80
8	1.43						
a	0.9708	0.9481	1.15	1.3054	1.7236	1.3676	0.8982
R	8	5	7	5	5	7	7
DF	5	2	4	2	2	3	4
X^2	1.16	2.39	0.50	1.03	0.33	0.20	1.44
P	0.95	0.30	0.97	0.60	0.85	0.98	0.84

	Column						
Rank	15	16	17	18	19	20	21
1	11.79	13.21	10.82	11.21	8.08	7.61	8.98
2	5.92	5.398	5.30	5.03	4.20	4.02	4.28
3	3.95	3.19	3.49	3.15	2.86	2.77	2.77
4	2.97	2.20	2.60	2.26	2.18	2.13	2.04
5	2.38	1.65	2.07	1.75	1.77	1.73	1.61
6	1.98	1.30	1.71	1.41	1.49	1.47	1.32
7		1.07		1.18	1.29	1.27	
8					1.13		
a	0.9946	1.2934	1.0291	1.1553	0.9438	0.9188	1.0687
R	6	7	6	7	8	7	6
DF	3	4	3	4	5	4	3
X^2	1.33	0.86	1.04	0.45	0.59	0.54	1.13
P	0.72	0.93	0.79	0.98	0.99	0.97	0.77

In Text 3, column 3, the distribution has been fitted by the Singh-Poisson distribution. This is, perhaps caused by the strong concentration of one of the POS.

Table 2.4.31
Rank-frequencies of POS in the Persian Text 4

	Column											
Rank	1	2	3	4	5	6	7	8	9	10	11	12
1	13	19	19	11	17	13	17	15	12	15	10	16
2	7	3	4	7	6	7	5	5	7	4	8	7
3	6	3	3	4	3	5	3	3	4	4	4	2
4	2	2	2	3	2	2	2	2	4	3	3	2
5	1	1	1	2	2	1	2	2	2	2	1	1
6	1	1	1	1		1	1	1	1	1	1	1
7		1		1		1		1		1	1	
8				1				1			1	
9											1	
Sum	30	30	30	30	30	30	30	30	30	30	30	29
WCS	.43	.63	.63	.37	.57	.43	.57	.50	.40	.50	.33	.55

	Column						
Rank	13	14	15	16	17	18	19
1	14	11	12	9	9	6	7
2	5	5	4	5	6	6	5
3	4	4	2	5	3	4	4
4	3	2	2	2	3	3	3
5	1	1	2	2	1	3	1
6	1	1	1	1	1	1	1
7		1	1	1			
8			1				
Sum	28	25	25	25	23	23	21
WCS	.50	.44	.48	.36	.39	.26	.33

Table 2.4.32
Fitting the zeta distribution to rank-frequencies of POS in Persian Text 4

	Column						
Rank	1	2	3	4	5	6	7
1	13.69	17.50	18.18	12.17	16.78	13.92	17.07
2	6.08	5.43	5.49	5.58	6.08	5.78	5.70
3	3.78	2.74	2.72	3.54	3.36	3.46	3.00
4	2.70	1.68	1.66	2.56	2.20	2.40	1.90
5	2.08	1.15	1.13	1.99	1.59	1.81	1.33
6	1.68	0.85	0.82	1.62		1.44	1.00
7		0.65		1.36		1.18	
8				1.17			

a	1.1722	1.6893	1.7281	1.1248	1.4649	1.2671	1.5836
R	6	7	6	8	5	7	6
DF	3	3	2	5	2	4	2
X^2	2.49	1.48	0.54	0.97	0.17	1.59	0.28
P	0.48	0.69	0.76	0.96	0.92	0.81	0.87

	Column						
Rank	8	9	10	11	12	13	14
1	14.47	12.50	14.20	11.9	16.48	13.78	11.38
2	5.57	6.12	5.77	5.40	5.51	5.60	4.83
3	3.19	4.03	3.41	3.42	2.90	3.30	2.92
4	2.15	3.00	2.35	2.47	1.84	2.27	2.05
5	1.58	2.38	1.76	1.92	1.30	1.70	1.55
6	1.23	1.97	1.38	1.57	0.97	1.34	1.24
7	0.99		1.13	1.32			1.02
8	0.84			1.13			
9				0.99			
a	1.3769	1.0303	1.2991	1.1267	1.5803	1.2997	1.2370
R	8	6	7	9	6	6	7
DF	4	3	4	5	2	3	4
X^2	0.27	1.02	1.03	2.47	0.74	0.82	0.66
P	0.99	0.80	0.91	0.78	0.69	0.84	0.96

	Column				
Rank	15	16	17	18	19
1	11.23	9.54	9.68	6.92	7.62
2	4.66	4.82	4.69	4.55	4.27
3	2.79	3.23	3.07	3.57	3.04
4	1.94	2.43	2.27	3.00	2.39
5	1.46	1.95	1.80	2.62	1.98
6	1.16	1.63	1.49	2.35	1.70
7	0.95	1.40			
8	0.80				
a	1.2680	0.9858	1.0456	0.6033	0.8374
R	8	7	6	6	6
DF	4	4	3	3	3
X^2	0.63	1.44	1.16	1.46	1.41
P	0.96	0.84	0.76	0.69	0.70

Table 2.4.33
Rank-frequencies of POS in the Persian Text 5

	Column								
Rank	1	2	3	4	5	6	7	8	9
1	17	17	16	14	12	13	7	13	6
2	4	3	6	6	5	5	5	6	6
3	4	3	4	5	3	2	4	4	5
4	2	2	2	3	2	2	3	3	3
5	2	2	1	1	2	2	3	1	3
6	1	1	1	1	2	2	3	1	3
7	1	1			1	2	2	1	2
8	1	1			1	1	1		
9					1		1		
10					1				
Sum	32	30	30	30	30	29	29	29	28
WCS	.53	.57	.53	.47	.4	.45	.24	.45	.21

	Column					
Rank	10	11	12	13	14	15
1	15	10	10	11	10	6
2	3	6	7	6	5	6
3	2	5	4	4	3	5
4	2	3	3	1	2	2
5	2	1	3	1	1	1
6	2	1		1	1	1
7	1	1			1	1
8	1				1	
Sum	28	27	27	24	24	22
WCS	.54	.37	.37	.46	.42	.27

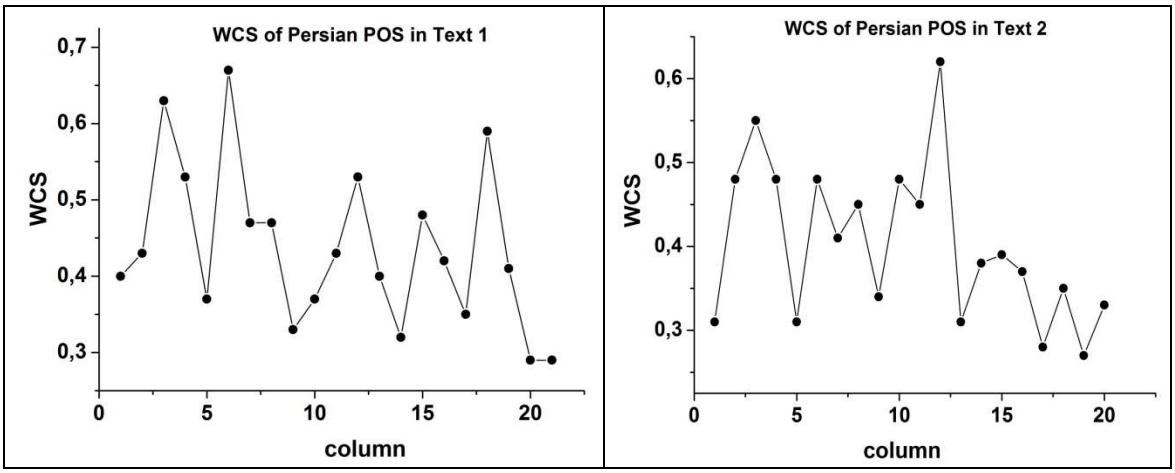
Table 2.4.34
Fitting the zeta distribution to rank-frequencies of POS in Persian Text 5

	Column						
Rank	1	2	3	4	5	6	7
1	15.99	15.98	16.29	14.18	12.59	12.59	7.48
2	5.91	5.47	5.82	6.05	5.41	5.41	4.69
3	3.31	2.92	3.18	3.67	3.30	3.30	3.57
4	2.19	1.87	2.08	2.58	2.33	2.33	2.94
5	1.59	1.33	1.49	1.96	1.77	1.77	2.53
6	1.22	1.00	1.14	1.57	1.42	1.42	2.24
7	0.98	0.79			1.18	1.18	2.02
8	0.81	0.64			1.00	1.00	1.84

9							1.70
a	1.4351	1.5468	1.4858	1.2300	1.2182	1.2182	0.6735
R	8	8	6	6	8	8	9
DF	4	3	3	3	4	4	6
X^2	1.02	1.62	0.40	1.23	1.18	1.18	1.13
P	0.91	0.66	0.94	0.75	0.88	0.88	0.98

	Column							
Rank	8	9	10	11	12	13	14	15
1	13.38	6.96	13.03	10.81	10.43	11.67	10.29	7.57
2	5.59	4.93	5.22	5.22	5.97	4.81	4.48	4.18
3	3.36	4.02	3.05	3.41	4.31	2.86	2.75	2.96
4	2.34	3.49	2.09	2.52	3.42	1.98	1.95	2.31
5	1.77	3.12	1.56	1.99	2.86	1.49	1.49	1.91
6	1.40	2.85	1.22	1.65		1.18	1.20	1.63
7	1.16	2.64	1.00	1.40			1.00	1.43
8			0.84				0.85	
a	1.2584	0.4984	1.3206	1.0502	0.8040	1.2789	1.2000	0.8561
R	7	7	8	7	5	6	8	7
DF	4	4	4	4	2	3	4	4
X^2	0.82	0.84	2.24	1.87	0.28	1.46	0.30	3.38
P	0.94	0.93	0.69	0.76	0.87	0.69	0.99	0.50

The WCS of Persian POS are presented in Figure 2.4.4



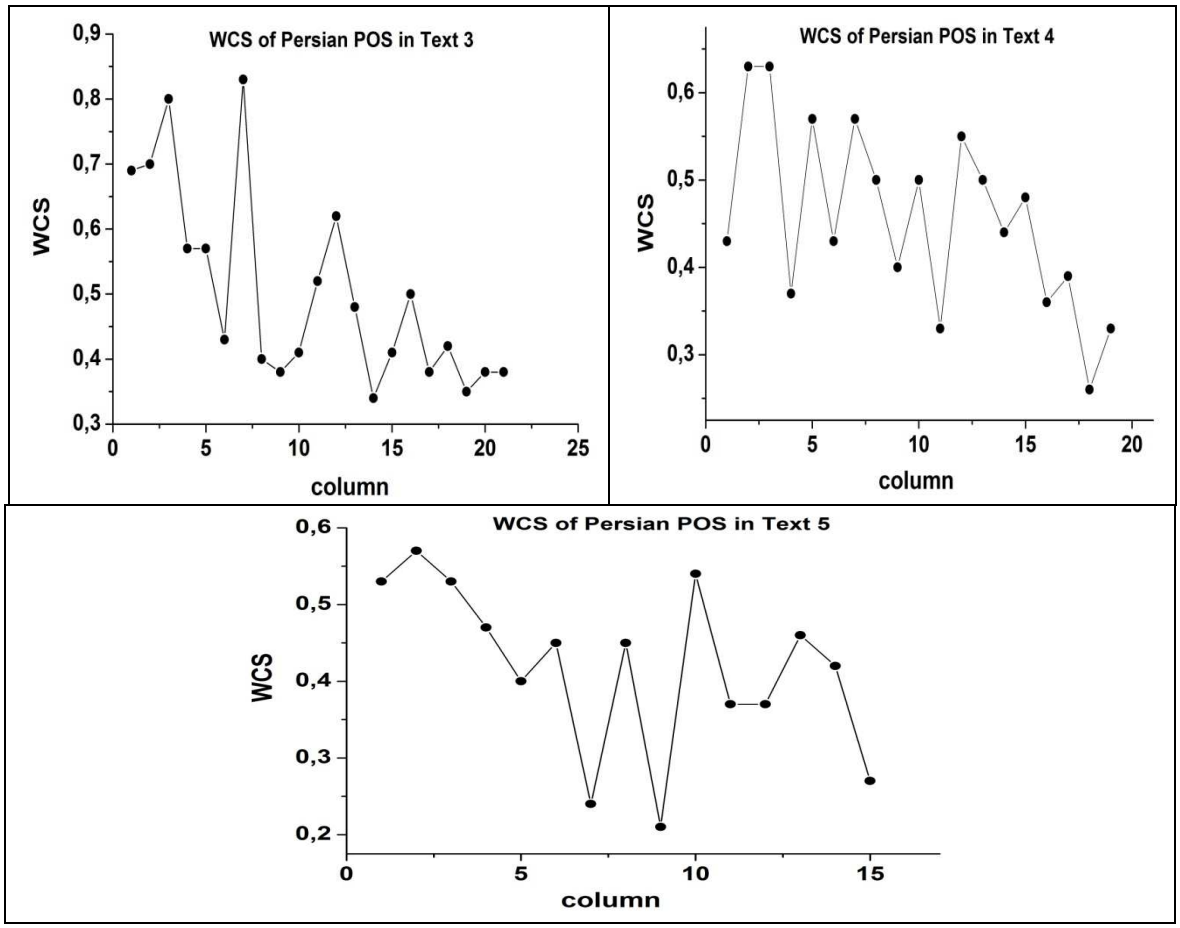


Figure 2.4.4. WCS of Persian POS

The rank-frequency distributions of parts of speech and the fitting for **French** are presented in Tables 2.4.35 to 2.4.44

Table 2.4.35
Rank-frequencies of POS in French Text 1

	Column								
Rank	1	2	3	4	5	6	7	8	9
1	14	18	14	15	15	16	19	10	13
2	9	7	14	10	8	7	6	8	7
3	9	7	7	8	7	5	4	7	7
4	5	7	4	5	6	5	4	7	6
5	5	4	4	4	5	5	3	4	4
6	3	3	4	4	3	5	3	3	2
7	3	2	2	2	3	3	2	2	2
8	2	1	1	1	1	2	2	1	1
9		1		1	1			1	
Sum	50	50	50	50	49	48	43	43	42
WCS	.28	.36	.28	.30	.31	.33	.44	.23	.31

	Column					
Rank	10	11	12	13	14	15
1	12	11	14	11	14	10
2	8	6	8	9	6	6
3	4	6	6	4	5	5
4	4	5	3	3	3	4
5	4	5	3	3	3	2
6	4	4	3	3	2	2
7	3	2	1	1	1	1
8	2	1				1
Sum	41	40	38	34	34	31
WCS	.29	.28	.37	.32	.41	.32

Table 2.4.36
Fitting the zeta distribution to rank-frequencies of POS in French, Text 1

	Column						
Rank	1	2	3	4	5	6	7
1	15.13	17.75	17.22	16.70	15.43	14.66	17.35
2	8.83	8.84	9.09	8.73	8.43	8.50	8.00
3	6.44	5.88	6.26	5.97	5.92	6.18	5.08
4	5.15	4.41	4.80	4.56	4.60	4.93	3.69
5	4.33	3.52	3.91	3.70	3.79	4.13	2.87
6	3.76	2.93	3.31	3.12	3.23	3.58	2.34
7	3.34	2.51	2.87	2.70	2.82	3.17	1.97
8	3.01	2.20	2.54	2.38	2.51	2.86	1.70
9		1.95		2.14	2.27		
a	0.7769	1.0049	0.9211	0.9359	0.8729	0.7864	1.1173
R	8	9	8	9	9	8	8
DF	5	6	5	6	6	5	5
X^2	1.74	3.41	4.81	2.95	2.69	1.62	1.16
P	0.88	0.76	0.44	0.82	0.85	0.90	0.95

	Column							
Rank	8	9	10	11	12	13	14	15
1	11.79	13.54	11.92	10.88	14.53	12.18	13.80	10.76
2	7.10	7.53	7.16	6.86	7.32	6.50	6.58	5.65
3	5.27	5.35	5.32	5.23	4.90	4.51	4.26	3.87
4	4.27	4.19	4.30	4.32	3.69	3.47	3.13	2.96
5	3.63	3.47	3.65	3.72	2.96	2.84	2.47	2.41
6	3.17	2.98	3.20	3.30	2.47	2.41	2.03	2.03
7	2.84	2.61	2.85	2.97	2.12	2.09	1.72	1.76

8 9	2.57 2.36	2.33	2.59	2.72				1.55
a	0.7324	0.8456	0.7349	0.6668	0.9890	0.9047	1.0696	0.9304
R	9	8	8	8	7	7	7	8
DF	6	5	5	5	4	4	4	5
X^2	4.73	2.66	0.82	2.33	1.16	1.92	0.60	1.36
P	0.58	0.75	0.96	0.80	0.88	0.75	0.96	0.93

Table 2.4.37
Rank-frequencies of POS in French, Text 2

	Column								
Rank	1	2	3	4	5	6	7	8	9
1	14	21	16	12	12	20	15	24	13
2	11	7	11	9	9	10	14	9	9
3	9	7	8	9	8	7	9	6	8
4	8	4	7	6	6	5	8	4	6
5	4	4	2	6	5	3	2	3	4
6	2	2	2	3	4	2	1	1	2
7	2	2	2	2	4	2		1	2
8		2	1	2	1				2
9		1							1
Sum	50	50	49	49	49	49	49	48	47
WCS	.28	.42	,33	.24	.24	.41	.31	.50	.28

	Column					
Rank	10	11	12	13	14	15
1	17	16	14	17	18	12
2	10	11	9	9	7	12
3	6	9	7	7	7	5
4	5	6	7	5	4	4
5	4	2	5	3	3	3
6	3	1	3	3	1	2
7	2	1			1	1
8					1	1
Sum	47	46	45	44	42	42
WCS	.36	.35	.31	.39	..43	.29

Table 2.4.38
Fitting the zeta distribution to rank-frequencies of POS in French, Text 2

	Column						
Rank	1	2	3	4	5	6	7
1	15.82	20.27	17.93	13.57	13.02	20.69	17.31
2	9.36	9.03	9.01	8.44	8.34	9.49	9.93
3	6.89	5.63	6.02	6.40	6.42	6.01	7.18
4	5.54	4.02	4.52	5.25	5.34	4.35	5.70
5	4.68	3.10	3.63	4.51	4.62	3.38	4.77
6	4.08	2.51	3.02	3.98	4.11	2.76	4.12
7	3.63	2.09	2.60	3.58	3.72	2.32	
8		1.79	2.27	3.27	3.42		
9		1.56					
a	0.7567	1.1664	0.9933	0.6848	0.6433	1.1250	0.8013
R	7	9	8	8	8	7	6
DF	4	6	5	5	5	4	3
X^2	4.12	1.41	4.58	3.31	2.37	0.61	7.33
P	0.39	0.97	0.47	0.65	0.80	0.96	0.06

	Column							
Rank	8	9	10	11	12	13	14	15
1	24.09	14.79	17.64	17.74	14.21	17.30	18.18	15.09
2	9.15	8.08	9.04	8.87	8.99	8.98	7.84	7.38
3	5.19	5.67	6.12	5.91	6.87	6.12	4.79	4.86
4	3.47	4.42	4.63	4.44	5.68	4.66	3.38	3.61
5	2.54	3.63	3.74	3.55	4.90	3.77	2.58	2.87
6	1.97	3.10	3.13	2.96	4.35	3.17	2.07	2.38
7	1.59	2.71	2.70	2.53			1.71	2.03
8		2.41					1.46	1.77
9		2.18						
a	1.3973	0.8720	0.9641	1.0001	0.6613	0.9462	1.2135	1.0308
R	7	9	7	7	6	6	8	8
DF	4	6	4	4	3	3	5	5
X^2	0.99	3.16	0.36	5.74	0.73	0.33	2.28	4.48
P	0.91	0.79	0.99	0.22	0.87	0.96	0.81	0.48

Table 2.4.39
Rank-frequencies of POS in French Text 3

	Column								
Rank	1	2	3	4	5	6	7	8	9
1	14	21	17	13	14	18	17	16	11
2	14	6	11	10	11	8	10	6	11
3	12	6	8	10	7	7	7	6	9
4	3	5	5	7	5	6	6	5	6
5	3	5	3	3	4	4	4	4	4
6	2	3	3	2	4	2	3	4	4
7	2	2	2	2	3	2	2	3	1
8		2	1	1	1	2		2	1
9				1				2	
Sum	50	50	50	49	49	49	49	48	47
WCS	.28	.42	.34	.27	.29	.37	.35	.33	.23

	Column					
Rank	10	11	12	13	14	15
1	20	14	17	10	12	15
2	5	8	6	10	10	8
3	5	6	6	8	6	7
4	4	5	5	6	6	3
5	4	5	3	4	3	2
6	3	4	3	1	2	1
7	3	1	1	1	1	1
8	2		1			
9						
Sum	46	43	42	40	40	37
WCS	.43	.33	.40	.25	.30	.41

Table 2.4.40
Fitting the zeta distribution to rank-frequencies of POS in French, Text 3

	Column						
Rank	1	2	3	4	5	6	7
1	17.67	19.25	18.82	15.75	15.64	18.00	17.74
2	9.55	9.25	9.23	8.47	8.77	9.01	9.39
3	6.66	6.03	6.08	5.90	6.25	6.01	6.47
4	5.16	4.45	4.52	4.56	4.92	4.51	4.97
5	4.23	3.52	3.60	3.73	4.08	3.61	4.05
6	3.60	2.90	2.98	3.17	3.50	3.01	3.42
7	3.14	2.46	2.54	2.76	3.08	2.58	2.97
8		2.14	2.22	2.45	2.76	2.26	

9				2.21			
a	0.8882	1.0562	1.0284	0.8942	0.8350	0.9981	0.9186
R	7	8	8	9	8	8	7
DF	4	5	5	6	5	5	4
X^2	9.51	2.10	2.06	7.23	2.02	1.31	0.70
P	0.05	0.84	0.84	0.30	0.85	0.93	0.95

	Column							
Rank	8	9	10	11	12	13	14	15
1	14.90	13.84	17.62	13.85	16.41	12.18	13.76	15.96
2	8.22	8.24	8.51	8.08	7.79	7.43	7.61	7.16
3	5.81	6.08	5.56	5.90	5.04	5.56	5.38	4.48
4	4.54	4.91	4.11	4.72	3.90	4.53	4.21	3.21
5	3.75	4.15	3.25	3.96	2.91	3.86	3.48	2.48
6	3.20	3.62	2.68	3.44	2.39	3.39	2.97	2.01
7	2.81	3.23	2.28	3.05	2.02	3.04	2.61	1.68
8	2.50	2.92	1.98		1.75			
9	2.26							
a	0.8579	0.7481	1.05	0.7771	1.0752	0.7135	0.8549	1.1563
R	9	8	8	7	8	7	7	7
DF	6	5	5	4	5	4	4	4
X^2	1.10	5.99	2.26	1.76	2.08	5.88	3.19	2.46
P	0.98	0.31	0.81	0.73	0.84	0.21	0.53	0.65

The first column of Text 3 is at the limit of significance but it is sufficient for our purposes. Better fitting can be obtained e.g. by the Hyperpoisson or the Conway-Maxwell-Poisson distributions.

Table 2.4.41
Rank-frequencies of POS in French, Text 4

	Column								
Rank	1	2	3	4	5	6	7	8	9
1	21	16	14	13	15	14	12	15	21
2	8	13	11	10	9	11	12	19	8
3	7	5	8	7	8	9	6	6	5
4	4	5	6	6	8	6	5	4	3
5	4	3	5	5	5	3	4	3	3
6	3	3	3	4	2	2	4	3	2
7	2	2	1	2		1	2	2	1
8	1	1	1				1	1	
9		1							
Sum	50	49	49	47	47	46	46	44	43
WCS	.42	.33	.28	.28	.32	.30	.26	.34	.49

	Column					
Rank	10	11	12	13	14	15
1	14	18	17	15	12	14
2	13	7	8	7	8	9
3	4	6	5	4	5	4
4	3	5	4	4	4	3
5	3	2	3	3	4	3
6	3	2	2	3	3	3
7	2	1	1	2	2	1
8	1			1	1	1
9						
Sum	43	41	40	39	39	38
WCS	.33	.44	.42	.38	.31	.37

Table 2.4.42
Fitting the zeta distribution to rank-frequencies of POS in French, Text 4

	Column						
Rank	1	2	3	4	5	6	7
1	20.65	18.68	16.08	14.08	15.04	16.19	14.55
2	9.31	8.78	8.82	8.70	9.41	8.78	8.21
3	5.85	5.64	6.21	6.56	7.15	6.14	5.88
4	4.20	4.12	4.84	5.37	5.88	4.76	4.64
5	3.25	3.23	3.99	4.60	5.06	3.91	3.86
6	2.64	2.65	3.41	4.05	4.47	3.33	3.32
7	2.21	2.24	2.98	3.64		2.90	2.92
8	1.89	1.94	2.66				2.62
9		1.70					
a	1.1487	1.0897	0.8655	0.6955	0.6771	0.8829	0.8249
R	8	9	8	7	6	7	8
DF	5	6	5	4	3	4	5
X^2	1.09	3.51	4.25	1.16	2.25	4.51	3.66
P	0.95	0.74	0.51	0.89	0.52	0.34	0.60

	Column							
Rank	8	9	10	11	12	13	14	15
1	16.41	20.87	16.54	17.82	17.25	14.63	12.68	14.80
2	8.11	8.24	7.96	7.94	7.74	7.19	7.01	7.04
3	5.37	4.79	5.19	4.94	4.85	4.75	4.95	4.56
4	4.01	3.26	3.83	3.54	3.48	3.54	3.87	3.35
5	3.20	2.41	3.02	2.72	2.69	2.82	3.20	2.64
6	2.66	1.89	2.50	2.20	2.18	2.34	2.74	2.17

7 8	2.27 1.98	1.54	2.12 1.84	1.84	1.82	1.99 1.74	2.40 2.14	1.84 1.59
a	1.0164	1.3401	1.0558	1.1667	1.1554	1.0239	0.8555	1.0717
R	8	7	8	7	7	8	8	8
DF	5	4	5	4	4	5	5	5
X^2	1.21	0.37	4.53	1.54	0.52	0.71	1.08	1.66
P	0.94	0.98	0.48	0.82	0.97	0.98	0.96	0.89

Table 2.4.43
Rank-frequencies of POS in French, Text 5

	Column								
Rank	1	2	3	4	5	6	7	8	9
1	18	24	15	16	15	14	13	22	17
2	10	6	10	9	12	11	11	7	14
3	9	6	8	7	7	8	9	7	7
4	7	4	8	6	6	7	5	4	5
5	3	3	6	5	5	6	4	3	2
6	2	3	2	3	2	3	3	2	1
7	1	2	1	3	2		2	2	
8		1		1	1				
9		1							
Sum	50	50	50	50	50	49	47	47	46
WCS	.36	.48	.30	.32	.30	.29	.28	.47	.37

	Column					
Rank	10	11	12	13	14	15
1	13	19	14	13	15	15
2	12	8	8	12	8	7
3	7	4	6	6	6	6
4	5	3	5	3	4	4
5	3	3	5	2	2	2
6	2	3	2	1	1	1
7	1	2	1	1	1	
Sum	43	42	41	38	37	35
WCS	.30	.45	.34	.34	.41	.43

Table 2.4.44
Fitting the zeta distribution to rank-frequencies of POS in French, Text 5

	Column						
Rank	1	2	3	4	5	6	7
1	18.92	22.41	15.67	16.34	17.23	14.97	15.18
2	9.63	9.08	9.34	9.00	9.10	9.73	8.84
3	6.48	5.35	6.91	6.34	6.25	7.56	6.44
4	4.90	3.68	5.57	4.95	4.80	6.32	5.15
5	3.94	2.75	4.72	4.09	3.91	5.50	4.32
6	3.30	2.17	4.12	3.49	3.30	4.91	3.75
7	2.84	1.78	3.67	3.06	2.87		3.33
8		1.49		2.73	2.54		
9		1.28					
a	0.9752	1.3031	0.7458	0.8612	0.9215	0.6219	0.7803
R	7	9	7	8	8	6	7
DF	4	6	4	5	5	3	4
X^2	3.87	1.86	4.69	1.66	3.61	1.12	2.57
P	0.42	0.93	0.3210	0.89	0.61	0.77	0.63

	Column							
Rank	8	9	10	11	12	13	14	15
1	21.28	19.56	15.78	18.24	14.25	15.99	15.90	15.26
2	9.08	9.38	8.25	8.13	7.81	7.36	7.16	7.12
3	5.52	6.10	5.65	5.07	5.49	4.67	4.49	4.56
4	3.87	4.49	4.31	3.62	4.28	3.38	3.23	3.32
5	2.95	3.55	3.50	2.79	3.53	2.64	2.50	2.60
6	2.35	2.92	2.95	2.26	3.01	2.15	2.02	2.13
7	1.95		2.56	1.89	2.63	1.81	1.69	
a	1.2287	1.0609	0.9354	1.1660	0.8679	1.1201	1.1506	1.0995
R	7	6	7	7	7	7	7	6
DF	4	3	4	4	4	4	4	3
X^2	0.96	4.75	3.95	0.63	2.15	5.04	1.74	1.34
P	0.92	0.19	0.41	0.97	0.71	0.28	0.78	0.72

The WCS of individual French texts are displayed graphically in Figure 2.4.5.

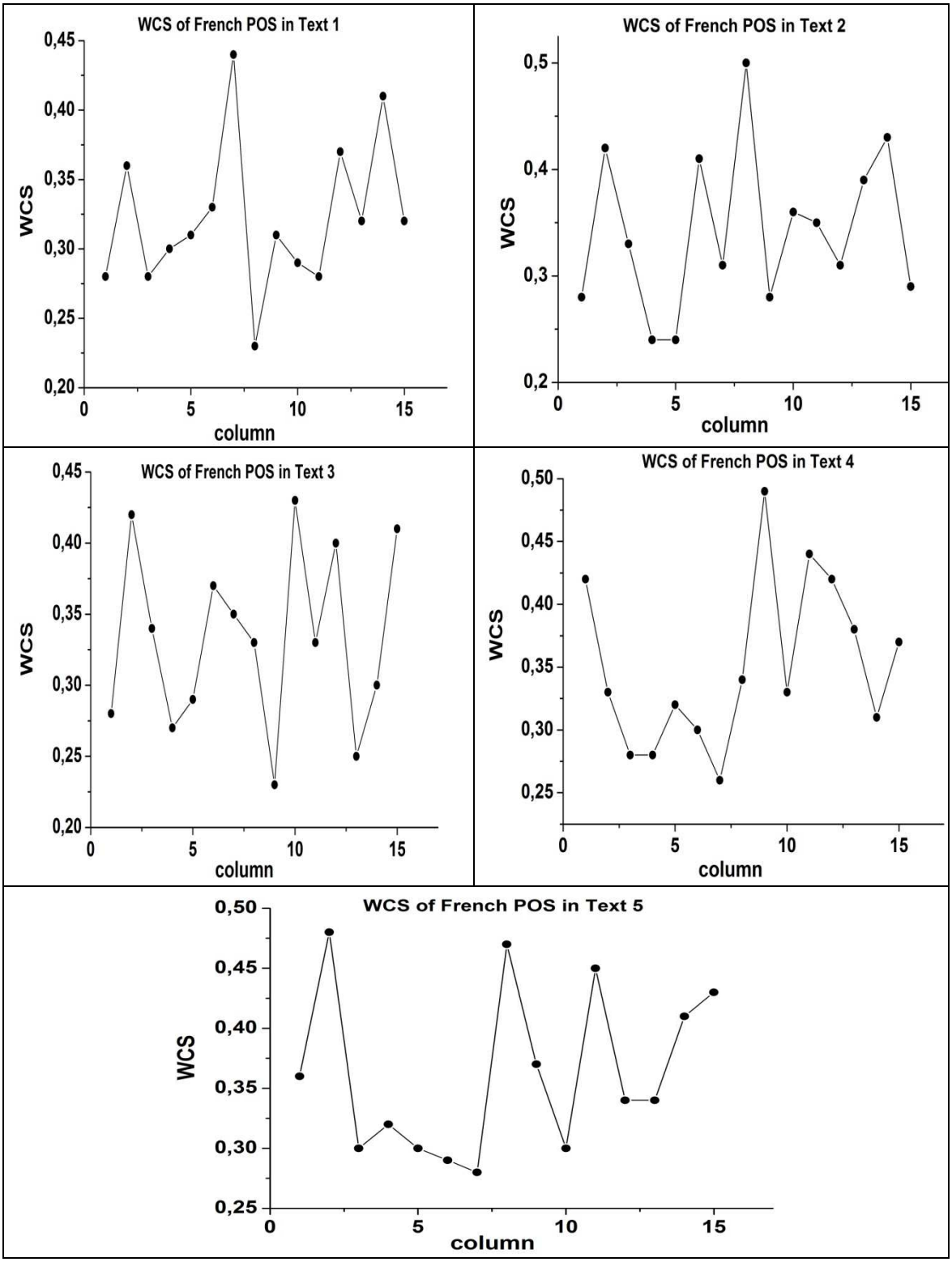


Figure 2.4.5. WCS of French POS

The distributions of **Turkish** parts of speech are presented in Tables 2.4.45 to 2.4.49

Table 2.4.45
The rank frequency of POS in Turkish: Text 1

	Column										
Rank	1	2	3	4	5	6	7	8	9	10	11
1	15	29	27	24	21	23	19	13	14	11	8
2	15	6	7	7	7	7	10	7	6	5	7
3	8	6	5	5	4	3	5	6	3	3	5
4	6	3	4	4	3	2	2	3	2	2	1
5	5	2	3	3	3	2	1	2	2	2	1
6	1	2	1	2	1	1		1	1	1	
7		1	1		1			1	1		
8					1			1			
SUM	50	49	48	45	41	38	37	34	29	24	22
WCS	.30	.59	.56	.53	.51	.61	.51	.38	.48	.46	.36

Table 2.4.46
The rank frequency of POS in Turkish: Text 2

	Column									
Rank	1	2	3	4	5	6	7	8	9	10
1	23	27	25	25	16	17	17	12	11	8
2	23	12	10	9	7	5	3	4	5	5
3	3	6	6	2	4	4	2	4	3	2
4	1	3	2	1	4	3	2	3	2	2
5		1	2	1	2	2	2	2	1	2
6		1	1	1	2		1	1	1	1
7			1		1		1			1
8							1			
SUM	50	50	47	39	36	31	29	26	23	21
WCS	.46	.54	.53	.64	.44	.55	.59	.46	.48	.38

Table 2.4.47
The rank frequency of POS in Turkish: Text 3

	Column						
Rank	1	2	3	4	5	6	7
1	30	29	32	25	35	23	26
2	8	12	10	14	3	10	11
3	5	3	3	3	3	4	4
4	3	2	2	2	3	2	2
5	2	1	1	2	2	2	1
6	2	1	1	2	2	2	
7		1	1	2	1	1	

8		1			1	1	
SUM	50	50	50	50	50	45	44
WCS	.60	.58	.64	.50	.70	.51	.59

	Column							
Rank	8	9	10	11	12	13	14	15
1	18	18	18	18	17	18	11	14
2	8	7	10	6	8	5	5	3
3	4	4	3	4	2	3	5	2
4	4	3	2	4	2	3	2	2
5	4	3	2	1	2	2	2	1
6	2	2	1	1	1	2	2	1
7		1	1	1	1		1	1
8			1					
SUM	40	38	38	35	33	33	28	24
WCS	.45	.47	.47	.51	.52	.55	.39	.58

Table 2.4.48
The rank frequency of POS in Turkish: Text 4

	Column						
Rank	1	2	3	4	5	6	7
1	33	27	25	32	24	29	25
2	10	10	10	8	11	5	11
3	5	3	6	3	4	4	4
4	2	3	4	2	3	3	3
5		3	3	1	2	3	2
6		3	1	1	2	2	1
7		1	1	1	1	1	1
8				1	1		
9				1			
SUM	50	50	50	50	48	47	47
WCS	.66	.54	.50	.64	.50	.62	.53

	Column							
Rank	8	9	10	11	12	13	14	15
1	24	28	21	22	19	17	8	11
2	12	5	11	6	4	6	6	4
3	3	3	3	4	4	5	5	3
4	3	3	3	3	2	1	3	2
5	2	3	1	1	2		1	1
6	2	2		1	1		1	1
7				1	1			1

8					1			
SUM	46	44	39	38	34	29	24	23
WCS	.52	.64	.54	.58	.56	.59	.33	.48

Table 2.4.49
The rank frequency of POS in Turkish: Text 5

	Column						
Rank	1	2	3	4	5	6	7
1	26	27	32	24	31	28	24
2	15	10	7	14	8	7	6
3	4	5	3	3	3	5	5
4	2	2	3	3	2	2	3
5	2	2	2	2	2	2	3
6	1	2	2	2	1	1	1
7		1	1	1	1		1
8		1		1			
SUM	50	50	50	50	48	45	43
WCS	.52	.54	.64	.48	.65	.62	.56

	Column							
Rank	8	9	10	11	12	13	14	15
1	19	23	20	17	14	14	12	12
2	11	10	6	9	6	4	6	3
3	4	3	3	3	5	3	3	3
4	3	2	2	2	2	3	2	1
5	2	1	2	2	1	1		1
6	2	1	2	1	1	1		1
7			1		1			
8			1					
SUM	41	40	37	34	30	26	23	21
WCS	.46	.58	.54	.50	.47	.54	.52	.57

The sequences of WCS are presented in Figure 2.4.6

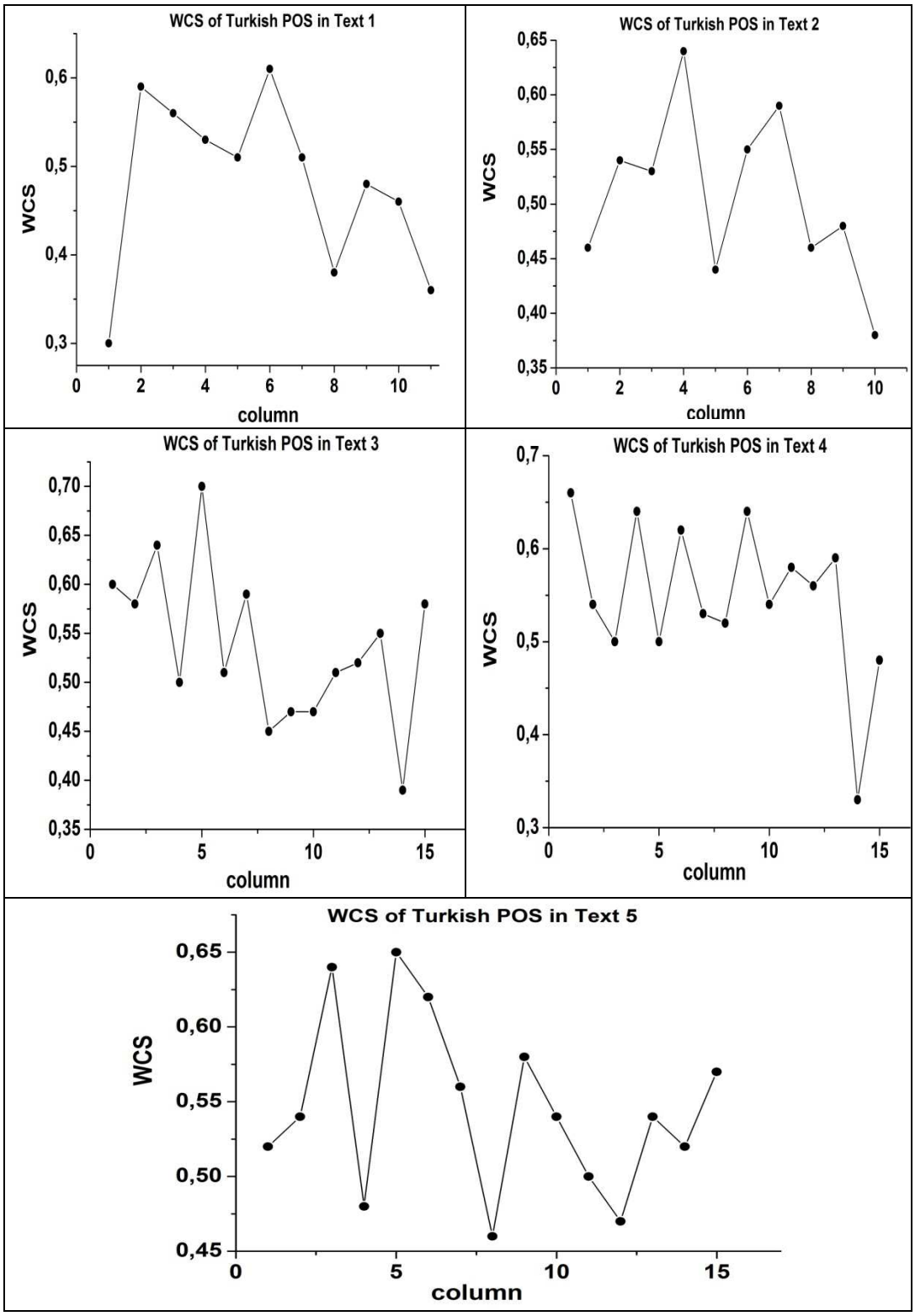


Figure 2.4.6. Sequences of WCS in Turkish texts

The fitting of the right truncated zeta distribution to these data is presented in Table 2.4.50 to 2.4.54. There is only one case which could be adjusted, namely

Text 2 column 1, because the first two values are equally high. This column can be fitted satisfactorily by the Poisson distribution. Evidently, here some boundary conditions are present which could be captured by means of other distributions; for example, it can be shown that the extended one-displaced positive Poisson distribution would yield $P = 0.53$ but we postpone this possibility until further texts have been analyzed.

Table 2.4.50
Fitting the right truncated zeta distribution to POS in Turkish: Text 1

	Column						
Rank	1	2	3	4	5	6	7
1	17.73	27.52	25.95	22.91	20.79	23.31	20.14
2	10.14	9.02	8.96	8.92	7.56	6.89	7.60
3	7.31	4.90	4.81	5.14	4.18	3.38	4.30
4	5.80	2.96	3.10	3.47	2.75	2.04	2.87
5	4.84	2.07	2.20	2.56	1.99	1.38	2.09
6	4.18	1.54	1.66	2.00	1.52	1.00	
7		1.20	1.31		1.21		
8					1.00		
a	0.8062	1.6089	1.5337	1.3610	1.4593	1.7574	1.4064
R	6	7	7	6	8	6	5
DF	3	4	4	3	4	2	2
X^2	5.25	1.63	1.37	0.62	0.79	0.21	1.77
P	0.15	0.80	0.85	0.89	0.94	0.90	0.41

	Column			
Rank	8	9	10	11
1	14.02	14.31	11.06	8.97
2	6.33	5.54	4.86	4.86
3	3.98	3.18	3.00	3.39
4	2.86	2.15	2.13	2.63
5	2.21	1.58	1.64	2.16
6	1.80	1.23	1.32	
7	1.50	1.00		
8	1.29			
a	1.14706	1.3676	1.1873	0.8862
R	8	7	6	5
DF	5	3	3	2
X^2	1.79	0.20	0.17	3.45
P	0.88	0.98	0.98	0.18

Table 2.4.51
Fitting the right truncated zeta distribution to POS in Turkish: Text 2

	Column						
Rank	1 Poisson	2	3	4	5	6	7
1	25.54	28.35	25.67	25.65	16.08	16.27	15.22
2	17.16	9.51	8.75	6.67	6.96	6.47	5.31
3	5.76	5.02	4.66	3.03	4.27	3.77	2.87
4	1.54	3.19	2.98	1.73	3.01	2.57	1.85
5		2.24	2.11	1.12	2.30	1.91	1.32
6		1.68	1.59	0.79	1.85		1.00
7			1.25		1.53		0.79
8							0.65
a	0.6717	1.5757	1.5533	1.9437	1.2076	1.3297	1.5196
R	-	6	7	6	7	5	8
DF	2	3	4	2	4	2	3
X^2	3.76	1.89	1.18	1.50	0.58	0.46	1.91
P	0.15	0.60	0.88	0.47	0.97	0.80	0.59

	Column		
Rank	8	9	10
1	11.48	11.28	8.31
2	5.29	4.60	4.06
3	3.36	2.72	2.67
4	2.43	1.88	1.98
5	1.90	1.41	1.57
6	1.55	1.11	1.30
7			1.11
a	1.1194	1.2939	1.0341
R	6	6	7
DF	3	3	4
X^2	0.79	0.21	0.60
P	0.85	0.98	0.96

Table 2.4.52
Fitting the right truncated zeta distribution to POS in Turkish: Text 3

	Column						
Rank	1	2	3	4	5	6	7
1	29.17	30.05	32.49	26.09	30.62	23.66	27.12
2	9.37	8.69	8.40	9.44	8.59	8.23	8.35
3	4.82	4.20	3.81	5.21	4.09	4.44	4.20
4	3.01	2.51	2.17	3.42	2.41	2.87	2.57

5	2.09	1.68	1.41	2.46	1.60	2.04	1.76
6	1.55	1.22	0.99	1.88	1.15	1.55	
7		0.92	0.73	1.50	0.86	1.22	
8		0.73			0.68	1.00	
a	1.6387	1.7902	1.9508	1.4665	1.8331	1.5227	1.6986
R	6	8	7	7	8	8	5
DF	3	4	3	4	4	4	2
X^2	0.37	2.14	0.66	4.03	5.57	0.86	1.35
P	0.95	0.71	0.88	0.40	0.23	0.93	0.51

	Column							
Rank	8	9	10	11	12	13	14	15
1	17.45	17.61	19.25	17.56	17.39	16.74	11.05	12.89
2	8.14	7.33	7.01	6.67	6.21	6.55	5.41	4.49
3	5.21	4.39	3.88	3.79	3.40	3.78	3.56	2.42
4	3.80	3.05	2.55	2.53	2.22	2.56	2.65	1.57
5	2.97	2.30	1.84	1.85	1.59	1.89	2.10	1.11
6	2.43	1.83	1.41	1.44	1.22	1.48	1.74	0.84
7		1.50	1.13	1.16	0.97		1.49	0.67
			0.93					
a	1.1005	1.2647	1.4575	1.3967	1.4851	1.3542	1.0309	1.5209
R	6	7	8	7	7	6	7	7
DF	3	4	4	4	3	3	4	3
X^2	0.74	0.46	1.81	1.49	1.24	0.89	0.97	0.95
P	0.86	0.98	0.77	0.83	0.74	0.83	0.91	0.81

Table 2.4.53
Fitting the right truncated zeta distribution to POS in Turkish: Text 4

	Column						
Rank	1	2	3	4	5	6	7
1	33.24	25.99	25.44	30.64	24.99	26.39	26.17
2	9.49	9.45	9.50	8.45	8.80	8.65	8.68
3	4.56	5.23	5.34	3.97	4.78	4.51	4.55
4	2.71	3.43	3.55	2.33	3.10	2.84	2.88
5		2.48	2.58	1.54	2.22	1.98	2.02
6		1.90	1.99	1.10	1.68	1.48	1.51
7		1.52	1.60	0.82	1.34	1.15	1.18
8				0.64	1.09		
9				0.52			
a	1.8086	1.4598	1.4211	1.8594	1.5054	1.6088	1.5916
R	4	7	7	9	8	7	7
DF	1	4	4	4	5	4	4
X^2	0.26	2.00	0.96	1.17	0.89	2.60	0.94
P	0.61	0.74	0.92	0.88	0.97	0.63	0.92

	Column							
Rank	8	9	10	11	12	13	14	15
1	24.85	25.18	22.02	21.54	17.75	16.69	8.87	10.86
2	8.94	8.33	7.86	6.97	6.23	6.33	4.88	4.43
3	4.92	4.36	4.30	3.60	3.38	3.59	3.44	2.62
4	3.22	2.76	2.80	2.25	2.19	2.40	2.69	1.80
5	2.31	1.93	2.01	1.57	1.56		2.22	1.35
6	1.77	1.44		1.16	1.19		1.90	1.07
7				0.91	0.94			0.87
					0.77			
a	1.4750	1.5961	1.4868	1.6284	1.5103	1.3992	0.8612	1.2952
R	6	6	5	7	8	4	6	7
DF	3	3	2	3	4	1	3	3
X^2	1.91	2.90	2.22	0.64	1.22	1.39	2.17	0.21
P	0.59	9.41	0.33	0.89	0.87	0.24	0.54	0.98

Table 2.4.54
Fitting the right truncated zeta distribution to POS in Turkish: Text 5

	Column						
Rank	1	2	3	4	5	6	7
1	28.24	27.32	30.20	25.91	30.44	27.45	22.88
2	9.53	9.05	8.87	9.18	8.23	8.20	8.07
3	5.05	4.74	4.33	5.00	3.83	4.04	4.38
4	3.22	3.00	2.60	3.25	2.23	2.45	2.85
5	2.27	2.10	1.76	2.33	1.46	1.66	2.04
6	1.70	1.57	1.27	1.77	1.04	1.21	1.55
7		1.23	0.97	1.41	0.77		1.23
8		0.99		1.15			
9							
a	1.5673	1.5942	1.7678	1.4969	1.8870	1.7438	1.5031
R	6	8	7	8	7	6	7
DF	3	4	3	5	3	3	4
X^2	4.32	0.59	1.26	3.71	0.44	0.60	1.37
P	0.23	0.96	0.74	0.59	0.93	0.90	0.85

	Column							
Rank	8	9	10	11	12	13	14	15
1	19.74	24.09	19.01	17.99	14.405	13.36	12.33	11.45
2	8.23	7.35	6.80	6.66	5.76	5.14	5.25	4.07
3	4.94	3.67	3.73	3.72	3.37	2.94	3.19	2.22

4 5 6 7	3.43 2.59 2.06	2.24 1.53 1.12	2.44 1.75 1.34 1.06 0.87	2.46 1.79 1.38	2.30 1.72 1.35 1.10	1.98 1.45 1.13	2.24	1.44 1.03 0.79
a R DF X^2 P	1.2616 6 3 1.33 0.72	1.7123 6 3 1.35 0.72	1.4818 8 4 0.74 0.95	1.4337 6 3 1.23 0.74	1.3217 7 4 1.25 0.87	1.3781 6 3 0.97 0.81	1.2313 4 1 0.15 0.70	1.4938 6 2 0.73 0.69

The fact that inspite of diverse allocation of POS at the individual positions in the sentence, the WCS has the typical structure of rank-frequencies, indicating that a kind of "vertical background mechanisms" in the spoken/written language might exist. It must be explainable on the basis of speaker/ hearer requirements in order to insert it in Köhler's control cycle but until now there are not enough text data available. The few so far studied examples do not allow generalizations.

2.5. Canonical syllable types

A sentence or verse can be transcribed as a string of canonical syllable types, e.g. V, CV, VC, CVC, …, where V and C represent the unique vowel and the consonants, respectively. This is rather a phonetic property. There is the possibility of obtaining also words having the form of a simple C, as is known from the Slavic languages. Either one considers them as parts of the next syllable, e.g. non-syllabic prepositions in Slavic languages may be considered as proclitics, or one analyzes each word separately. For each sentence one obtains a string of symbols. As a matter of fact, one can consider syllable types as elements of a two-dimensional vectorspace, whose components are the number of consonants before and behind the vowel. However, we will not pursue this approach here (cf. Zörnig, Altmann 1993; Obradović et al. 2010; Kelih, Mačutek 2013).

Let us first consider the canonical syllable types in the German poem by J.W.v. Goethe, *Der Erlkönig*. The syllable-type frequencies are presented in Table 2.5.1.

Table 2.5.1
Syllable type frequencies in *Der Erlkönig*

Syllable	1	2	3	4	5	6	7	8	9
V	0	0	0	0	0	0	2	0	0
CV	5	19	5	7	8	8	8	5	9
VC	9	0	3	2	3	4	4	2	2

CCV	0	1	0	5	0	0	0	0	0
CVC	11	6	23	12	19	13	9	16	8
VCC	4	2	0	2	0	1	4	2	2
CVCC	2	4	1	2	1	4	6	4	7
CCVC	0	0	0	1	1	2	0	2	0
CVCCC	1	0	0	0	0	0	0	0	0
CCVCC	0	0	0	0	0	0	0	0	0
CCCVC	0	0	0	0	0	0	0	1	0
CCCVCC	0	0	0	0	0	0	0	0	1
CVCCCC	0	0	0	1	0	0	0	0	0
Sum	32	32	32	32	32	32	32	32	29
WCS	0.34	0.59	0.72	0.38	0.59	0.41	0.28	0.50	0.28

Again, we rank the distributions in each column and try to capture it by the Hyperpoisson distribution. The results are presented in Table 2.5.2

Table 2.5.2
Fitting the Hyperpoisson distribution to ranked syllable types in the poem *Der Erlkönig* by J. Goethe

Rank	1	2	3	4	5
1	10.78	17.07	20.45	10.31	18.93
2	8.82	8.32	7.85	7.43	7.92
3	5.97	3.83	2.64	5.16	3.19
4	3.45	1.67	1.06	3.46	1.24
5	1.74	1.12		2.24	0.72
6	1.25			1.41	
7				0.85	
8				1.13	
a	3.9281	8.1938	2.6863	19.4098	10.7206
b	4.8010	16.8188	7.0004	26.9444	25.6103
DF	3	2	1	4	1
X^2	0.35	0.95	1.41	1.20	0.01
P	0.95	0.62	0.24	0.88	0.91

Rank	6	7	8	9
1	12.65	8.74	12.20	9.00
2	8.30	8.11	7.76	8.28
3	5.08	6.40	4.84	5.82
4	2.91	4.39	2.97	3.32
5	1.57	2.66	1.78	1.59
6	1.48	2.70	1.05	1.00
7			1.39	
a	9.0333	5.2650	33.6630	2.9941

b	13.7667	5.6737	52.9474	3.2556
DF	3	3	4	2
X^2	0.94	0.92	3.61	0.84
P	0.82	0.82	0.46	0.66

As can be seen, the fitting is satisfactory in each case. The positions 10 to 12 have a smaller number of cases, hence we omitted them. Writing the parameters *a* and *b* for each text and ordering them according to increasing *a*, we obtain a very clear relation presented in Table 2.5.3 and Figure 2.5.2.

Table 2.5.3
The dependence of parameter *b* on *a*

a	b	b computed
2.6863	7.0004	4.9906
2.9941	3.2556	5.5194
3.9281	4.801	7.1020
5.265	5.6737	9.3218
8.1938	16.8188	14.0557
9.0333	13.7667	15.3881
10.7206	25.6103	18.0401
19.4098	26.9444	31.3047
33.663	52.9474	52.1967
k = 1.9938; c = 0.9285; R^2 = 0.9356		

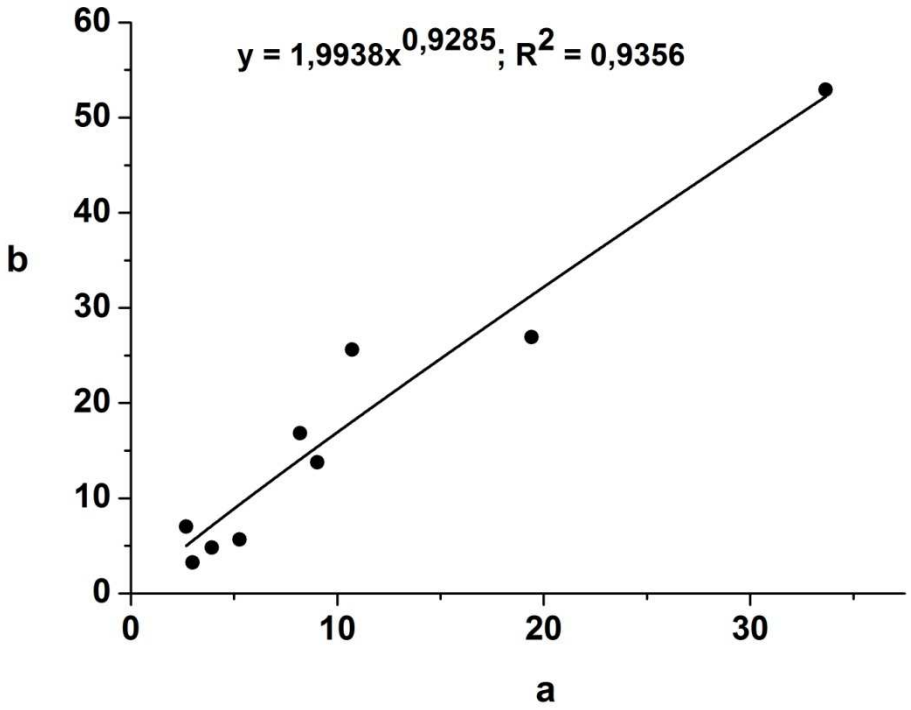


Figure 2.5.2. $b = f(a)$ for syllable types in German

Again, the function relating a and b is a power function, here having the form $b = 1.9938a^{0.9285}$ with $R^2 = 0.9356$.

The last analysis shows that whatever the text unit might be – from phonemics to semantics, lexicology, etc. – all of them are vertically structured. This is a hint at the fact that though sentence is a one-dimensional phenomenon, text may be a two-dimensional structure. New sentences bear the traces of the preceding ones. In poetry, the structuration is more deterministic, but as shown in all cases above, in prose there is something that cannot be considered chaotic. Some stochastic laws control the structure of the text which is presented as a sequence of syllables, words, sentences, etc.

3. Comparison of Consensus Strings

One can consider the elements of a WCS as a set of numbers and analyze them by proposing various indicators known from exploratory statistics. In this case we may speak about the static perspective. Or one can consider the WCS as a sequence of numbers and analyze their course. In that case we may speak about the dynamic perspective.

3.1. Static approach

Weighted consensus strings have their specific shape according to language, text type and considered entity. As can be seen, there is generally a strong oscillation which differs considerably depending on text and entity. In principle one could apply a Fourier analysis but due to the irregularities one cannot expect any reasonable result. Nevertheless, we can compare the WCS's restricting us to a given language, given text type and given entity. One can propose a number of different indicators, expressing important characteristics of a WCS. In this section we study first some static indicators. The comparisons may be performed as follows. A WCS is a sequence of numbers $(x_1,..,x_N)$ over the interval [0,1].

(1) The first, simple possibility is the consideration of the *mean value*

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \ .$$

The greater is the mean, the more structured or more stereotyped is the text from the point of view of the given entity. A high value of an element of WCS means a concentration of the same entity at this position. A low value means a low concentration, i.e. no structuring trend at the given position. But computing simply the mean may yield similar numbers both to texts with strong oscillation and with small oscillation. Thus the mean itself does not express the shape of the WCS but the kind of the steady state.

(2) Another important characteristic is the standard deviation. But high values of this indicator indicate merely large differences between the columns which may or may not be due to a strong oscillation. This indicator is not adequate for measuring oscillation

(3) A apparently good criterion allowing us to distinguish languages and entities is the use of Ord's criterion. We compute for each text and the given entity two functions made up of the moments of the distribution, namely

$$m_1' = \frac{1}{N}\sum_{x=x_{\min}}^{x_{\max}} x f_x \ , \qquad \text{(mean, average)}$$

$$m_2 = \frac{1}{N} \sum_{x=x_{\min}}^{x_{\max}} (x - m_1')^2 f_x \qquad \text{(variance, second central moment)}$$

$$m_3 = \frac{1}{N} \sum_{x=x_{\min}}^{x_{\max}} (x - m_1')^3 f_x \qquad \text{(asymmetry, third central moment)}$$

and set up the indicators:

$$I = \frac{m_2}{m_1'}, \qquad S = \frac{m_3}{m_2}.$$

Then we represent the pairs <*I*, *S*> obtained from the individual texts as points in the Cartesian coordinate system. Table 3.1.1 presents the values of *I* and *S* for several texts.

Table 3.1.1
Ord's indicators of WCS. Rank-frequencies of POS

Text	**I**	**S**
Chinese Text 1	3.1234	-0.2017
Chinese Text 2	3.1828	-0.0644
Chinese Text 3	3.1148	0.0370
Chinese Text 4	3.2961	0.4163
Chinese Text 5	3.3298	0.0214
French Text 1	2.3163	-0.1135
French Text 2	2.2400	-0.3178
French Text 3	2.3410	-0.0463
French Text 4	2.2873	-0.6693
French Text 5	2.4064	-0.3207
Persian Text 1	3.3939	0.6449
Persian Text 2	3.0932	0.6732
Persian Text 3	3.7692	1.4324
Persian Text 4	3.0442	0.6900
Persian Text 5	2.6162	0.5202
Polish Text 1	2.1421	-0.3501
Polish Text 2	2.3149	0.1739
Polish Text 3	2.3410	-0.0463
Polish Text 4	2.2920	-0.6521
Polish Text 5	2.1485	-0.2518

Turkish Text 1	1.5239	0.3428
Turkish Text 2	1.423	0.1652
Turkish Text 3	2.4807	0.4901
Turkish Text 4	2.3346	0.1344
Turkish Text 5	2.3569	0.2970

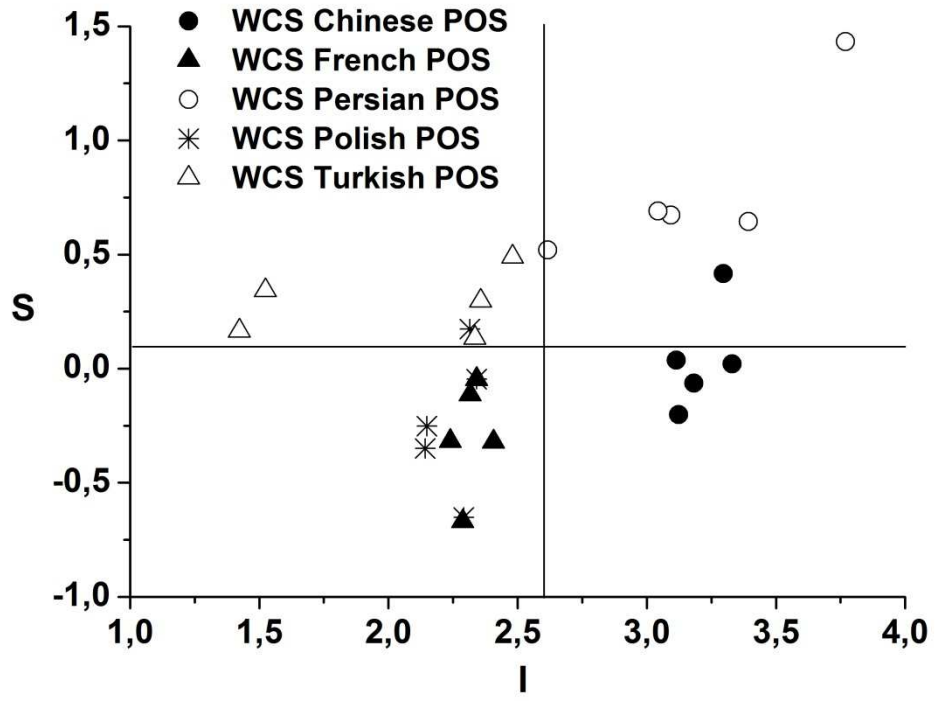


Figure 3.1.1. Ord's indicator of WCS of POS in five languages

The two orthogonal lines $I = 2.6$ and $S = 0.1$ in Fig. 3.1.1 illustrate a first attempt to distinguish the regions in which the points of the considered five languages are located. These lines divide the *IS* plane into the four quadrants

$$Q_1 := \{(I,S) \mid I \le 2.6, S \ge 0.1\},\ Q_2 := \{(I,S) \mid I \ge 2.6, S \ge 0.1\},$$
$$Q_3 := \{(I,S) \mid I \le 2.6, S \le 0.1\},\ Q_4 := \{(I,S) \mid I \ge 2.6, S \le 0.1\}.$$

Disregarding a few exceptions one can see that Q_1 contains Turkish texts, Q_2 Persian texts, Q_3 contains French and Polish texts and Q_4 is the region of Chinese texts. The analysis of further texts will show whether this initial subdivision of the IS-plane can be validated.

For the WCS of word lengths in the five languages we obtain the results presented in Table 3.1.2 and Figure 3.1.2.

Table 3.1.2
Ord's indicators of WCS of word length

Text	**I**	**S**
French Text 1	3.1563	-0.2414
French Text 2	3.1648	0.0818
French Text 3	3.3564	-0.1767
French Text 4	3.2895	0.2683
French Text 5	3.2170	-0.0738
Persian Text 1	3.3775	-0.0571
Persian Text 2	3.0952	-0.5411
Persian Text 3	3.5010	0.4378
Persian Text 4	2.8794	-0.1984
Persian Text 5	2.0183	0.0949
Polish Text 1	3.5010	0.4378
Polish Text 2	2.4678	-0.0775
Polish Text 3	2.6677	0.0351
Polish Text 4	3.5741	0.6912
Polish Text 5	2.5361	0.0917
Turkish Text 1	2.0954	-0.3342
Turkish Text 2	1.7291	0.1421
Turkish Text 3	2.2599	-0.0351
Turkish Text 4	2.2142	-0.0580
Turkish Text 5	3.6174	-0.0328

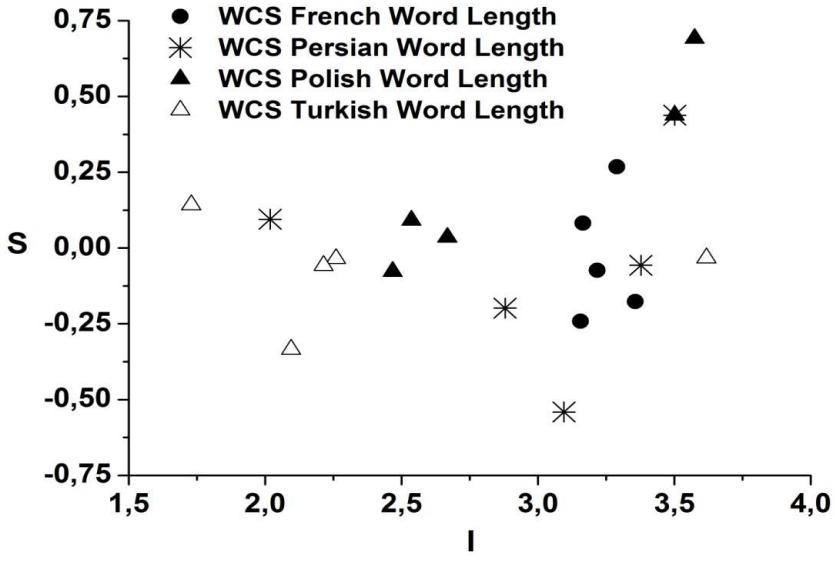


Figure 3.1.2. Ord's indicator of WCS of lengths in four languages

Here, the identification of regions corresponding to languages is more complex.

3.2. Dynamic approach

At first glance the arc length seems to be an adequate indicator to compare WCS´s. However, the use of the arc measure by means of the Euclidean distance is not suitable since the components of the WCS are all smaller than 1 while the vertical component in the graphical illustration of the arc is exactly 1 so that any part of the arc has a length slightly greater than 1, e.g. 0.44 against 0.55 yields $[(0.44 - 0.55)^2 + 1]^{1/2} = 1.0060$. For this reason we use another indicator: Let the number of extremes (minima and maxima) be *m* and denote the number of all points (= columns) by *n*. Since the first and the last points are always extremes, one can omit them and define the *non-smoothness indicator* (cf. Popescu et al. 2010: 95 ff.) as

$$NS = \frac{m-2}{n-2}.$$

Since this a simple proportion, its variance is

$$Var(NS) = \frac{NS(1-NS)}{n-2}.$$

The comparison of two texts may be performed either by applying the binomial distribution or asymptotically the normal distribution, hence one obtains

$$u = \frac{|NS_1 - NS_2|}{\sqrt{Var(NS_1) + Var(NS_2)}}.$$

Consider, e.g. the Persian texts 1 and 2 where one obtains $NS_1 = 0.65$, $n_1 = 22$ and $NS_2 = 0.6842$, $n_2 = 21$. For the difference we obtain

$\mathrm{Var}(NS_1) = 0.65(1 - 0.65)/(22\text{-}2) = 0.0114,$
$\mathrm{Var}(NS_2) = 0.6842(1 - 0.6842)/(21\text{-}2) = 0.0114,$ hence

$u(T1,T2) = |0.65 - 0.6842|/(0.0114 + 0.0114)^{1/2} = 0.23$

which is a non-significant difference. But for Persian texts 2 and 3 we obtain $u = 2.17$, a significant difference. Comparing all texts in one language with all the other ones, one can compute the extent of homogeneity in the given text type for the given language. In order to obtain some unique measure, one can take the means of all *u*.

For our data we obtain the results concerning *word lengths* presented in Tables 3.2.1 to 3.2.8.

Table 3.2.1
Non-smoothness of WCS of word lengths in Persian texts

Text	**n**	**m**	**NS**	**Var(NS)**
Persian WL Text 1	22	15	0.6500	0.0114
Persian WL Text 2	21	15	0.6842	0.0114
Persian WL Text 3	21	13	0.5789	0.0128
Persian WL Text 4	19	11	0.5294	0.0147
Persian WL Text 5	14	7	0.4167	0.0203

Table 3.2.2
Non-smoothness of WCS of word lengths in Turkish texts

Text	**n**	**m**	**NS**	**Var(NS)**
Turkish WL Text 1	12	8	0.6000	0.0240
Turkish WL Text 2	12	8	0.6000	0.0240
Turkish WL Text 3	15	12	0.7692	0.0137
Turkish WL Text 4	15	9	0.5385	0.0191
Turkish WL Text 5	23	14	0.5714	0.0117

Table 3.2.3
Non-smoothness of WCS of word lengths in Polish texts

Text	**n**	**m**	**NS**	**Var(NS)**
Polish WL Text 1	21	11	0.4737	0.0131
Polish WL Text 2	16	9	0.5000	0.0179
Polish WL Text 3	17	11	0.6000	0.0160
Polish WL Text 4	21	13	0.5789	0.0128
Polish WL Text 5	16	11	0.6429	0.0164

Table 3.2.4
Non-smoothness of WCS of word lengths in French texts

Text	**n**	**m**	**NS**	**Var(NS)**
French WL Text 1	20	16	0.7778	0.0096
French WL Text 2	20	14	0.6667	0.0123
French WL Text 3	20	14	0.6667	0.0123
French WL Text 4	20	15	0.7222	0.0111
French WL Text 5	20	13	0.6111	0.0132

Table 3.2.5
Differences in non-smoothness of WCS of word lengths in Persian texts

Text	1	2	3	4	5
1	0.0000	0.9832	1.1265	0.7474	0.3749
2	0.9832	0.0000	2.1354	1.7000	1.2105
3	1.1265	2.1354	0.0000	0.3344	0.6002
4	0.7474	1.7000	0.3344	0.0000	0.2886
5	0.3749	1.2105	0.6002	0.2886	0.0000

Table 3.2.6
Differences in non-smoothness of WCS of word lengths in Polish texts

Text	1	2	3	4	5
1	0.0000	0.1494	0.7404	0.6537	0.9851
2	0.1494	0.0000	0.5431	0.4503	0.7716
3	0.7404	0.5431	0.0000	0.1243	0.2383
4	0.6537	0.4503	0.1243	0.0000	0.3745
5	0.9851	0.7716	0.2383	0.3745	0.0000

Table 3.2.7
Differences in non-smoothness of WCS of word lengths in French texts

Text	1	2	3	4	5
1	0.0000	0.7507	0.7507	0.3864	1.1040
2	0.7507	0.0000	0.0000	0.3628	0.3482
3	0.7507	0.0000	0.0000	0.3628	0.3482
4	0.3864	0.3628	0.3628	0.0000	0.7127
5	1.1040	0.3482	0.3482	0.7127	0.0000

Table 3.2.8
Differences in non-smoothness of WCS of word lengths in Turkish texts

Text	1	2	3	4	5
1	0.0000	0.0000	0.8721	0.2964	0.1513
2	0.0000	0.0000	0.8721	0.2964	0.1513
3	0.8721	0.8721	0.0000	1.2748	1.2432
4	0.2964	0.2964	1.2748	0.0000	0.1879
5	0.1513	0.1513	1.2432	0.1879	0.0000

The means of u in individual languages are

Polish	1.0062
French	1.0253
Turkish	1.5770
Persian	1.9002

The results concerning POS are presented in Tables 3.2.9 to 3.2.13

Table 3.2.9
Differences in non-smoothness of WCS of POS in Chinese texts

Text	1	2	3	4	5
1	0.0000	0.3482	0.3384	0.0000	0.6749
2	0.3482	0.0000	0.6890	0.3482	1.0299
3	0.3384	0.6890	0.0000	0.3384	0.3347
4	0.0000	0.3482	0.3384	0.0000	0.6749
5	0.6749	1.0299	0.3347	0.6749	0.0000

Table 3.2.10
Differences in non-smoothness of WCS of POS in French texts

Text	1	2	3	4	5
1	0.0000	0.0000	0.4134	0.8163	0.4134
2	0.0000	0.0000	0.4134	0.8163	0.4134
3	0.4134	0.4134	0.0000	0.3982	0.0000
4	0.8163	0.8163	0.3982	0.0000	0.3982
5	0.4134	0.4134	0.0000	0.3982	0.0000

Table 3.2.11
Differences in non-smoothness of WCS of POS in Persian texts

Text	1	2	3	4	5
1	0.0000	2.0184	0.3250	2.3801	1.2727
2	2.0184	0.0000	1.6693	0.3397	0.5302
3	0.3250	1.6693	0.0000	2.0222	0.9665
4	2.3801	0.3397	2.0222	0.0000	0.8314
5	1.2727	0.5302	0.9665	0.8314	0.0000

Table 3.2.12
Differences in non-smoothness of WCS of POS in Polish texts

Text	1	2	3	4	5
1	0.0000	0.5002	1.3744	1.8038	0.0000
2	0.5002	0.0000	0.8611	1.2738	0.5002
3	1.3744	0.8611	0.0000	0.3982	1.3744
4	1.8038	1.2738	0.3982	0.0000	1.8038
5	0.0000	0.5002	1.3744	1.8038	0.0000

Table 3.2.13
Differences in non-smoothness of WCS of POS in Turkish texts

Text	1	2	3	4	5
1	0.0000	1.5755	0.6532	1.5027	0.2800
2	1.5755	0.0000	1.0531	0.1871	1.4535
3	0.6532	1.0531	0.0000	0.9472	0.4134
4	1.5027	0.1871	0.9472	0.0000	1.3744
5	0.2800	1.4535	0.4134	1.3744	0.0000

The mean u for the comparison of individual texts in individual languages are

French	0.4083
Chinese	0.4777
Turkish	0.9918
Polish	0.9899
Persian	2.2356

3.3. Further possible indicators

We finally indicate possible indicators which might be used in future studies. Instead of the number of oscillations one could evaluate their "intensity". A possible formalization could be the range $R = \frac{Max(x_1,...,x_m) - Min(x_1,...,x_m)}{m}$, where a WCS is again denoted by $(x_1,\ldots,x_m)$. Furthermore, smoothing techniques of time series analysis could be applied. In particular, one can substitute a given WCS by a corresponding sequence of moving averages. It might be easier to compare "smoothed versions" of the WCS´s than the original ones.

Furthermore von Neumanns´s (1941) trend indicator might be a useful statistic. In the remainder of the section we restrict ourselves to the latter alternative.

This indicator calculates the "mean square successive differences"

$$D = \frac{1}{m-1}\sum_{i=1}^{m-1}(x_i - x_{i+1})^2$$

which will be divided by the sample variance. The statistic might be adequate to express the "amount of oscillation" more precisely because it takes the "stepwise change" of the WCS into account.

For example, the syllable types in *Der Erlkönig* by Goethe yield the successive squared differences:

[0.0625, 0.0169, 0.1156, 0.0441, 0.0324, 0.0169, 0.0484, 0.0484,]

whose sum, divided by $m - 1 = 8$ yields $D = 0.04815$.

Usually, one normalizes *D* by dividing it by the variance of the values in WCS. One obtains the indicator *DN*, representing *D*/Var(WCS).

For the word lengths in individual texts and languages we obtain the results presented in Table 3.3.1

Table 3.3.1
Mean square successive difference for word lengths

French WL	**m**	**D**	**Var(WCS)**	**DN**
T1	20	0.0289	0.009371	3.0867
T2	20	0.0111	0.008442	1.3167
T3	20	0.0050	0.004550	1.1000
T4	20	0.0140	0.006237	2.2464
T5	20	0.0113	0.007599	1.4933
Persian WL	**m**	**D**	**Var(WCS)**	**DN**
T1	22	0.0058	0.002472	2.3483
T2	21	0.0100	0.004063	2.4610
T3	21	0.0033	0.002289	1.4373
T4	19	0.0093	0.004623	2.0190
T5	14	0.0331	0.011541	2.8713
Polish WL	**m**	**D**	**Var(WCS)**	**DN**
T1	21	0.0033	0.002289	1.4373
T2	16	0.0045	0.001878	2.4028
T3	17	0.0038	0.001790	2.1119
T4	21	0.0051	0.003475	1.4533
T5	16	0.0032	0.001306	2.4191
Turkish WL	**m**	**D**	**Var(WCS)**	**DN**

T1	12	0.0138	0.011202	1.2295
T2	11	0.0040	0.002289	1.7649
T3	15	0.0057	0.002331	2.4234
T4	15	0.0015	0.000852	1.7598
T5	23	0.0049	0.002099	2.3320

For POS, the results are presented in Table 3.3.2.

Table 3.3.2
Mean square successive difference for POS

Chinese POS	**m**	**D**	**Var(WCS)**	**DN**
T1	20	0.0116	0.004609	2.5072
T2	20	0.0035	0.002158	1.6195
T3	20	0.0055	0.003269	1.6681
T4	20	0.0055	0.003318	1.6698
T5	20	0.0077	0.004282	1.8017
French POS	**m**	**D**	**Var(WCS)**	**DN**
T1	15	0.0074	0.002931	2.5244
T2	15	0.0139	0.005678	2.4518
T3	15	0.0100	0.004024	2.4941
T4	15	0.0065	0.004470	1.4591
T5	15	0.0099	0.004750	2.0859
Persian POS	**m**	**D**	**Var(WCS)**	**DN**
T1	21	0.0202	0.011251	1.7913
T2	20	0.0152	0.008922	1.7048
T3	21	0.0262	0.022151	1.1848
T4	19	0.0185	0.010980	1.6889
T5	15	0.0236	0.012241	1.9256
Polish POS	**m**	**D**	**Var(WCS)**	**DN**
T1	15	0.0219	0.008269	2.6469
T2	15	0.0209	0.009378	2.2332
T3	15	0.0100	0.004024	2.4941
T4	15	0.0065	0.004374	1.4778
T5	15	0.0244	0.009598	2.5407
Turkish POS	**m**	**D**	**Var(WCS)**	**DN**
T1	11	0.0144	0.009689	1.4821
T2	10	0.0111	0.005979	1.8510
T3	15	0.0136	0.006364	2.1371

T4	15	0.0139	0.006827	2.0424
T5	15	0.0076	0.003440	2.1948

It would be premature to draw conclusions based on only five texts in one language but it is to be expected that the analysis of extensive text material would enable us to classify text types, set up a typology of languages, to study the evolution of text types etc. We merely prepared a way to a new view.

In Popescu et al. (2010) another indicator of *roughness* has been defined. It consists of the above NS indicator which is multiplied by the relative arc length between positions. The arc length is defined as

$$L = \sum_{i=1}^{m-1} \left[(x_i - x_{i+1})^2 + 1^2 \right]^{1/2}$$

But in our case the minimum value of x_i is $1/n_i$ where n_i is the column sum (not 0) and the maximum value is 1, hence our indicator will be different.

It is sufficient to relativize the D by dividing it by its maximum, that is,

$$D_{rel} = \frac{\sum_{i=1}^{m-1} (n_i - n_{i+1})^2}{\sum_{i=1}^{m-1} (1/n_i - 1)^2}.$$

Other indicators can be proposed but it is better to wait until more texts are analyzed.

4. Conclusions

The appearance of regularities in the WCS, in the rank-frequency distribution of columns as well as in the distribution of numerical values in the columns (representing the classes) is a sufficient reason to conjecture that the text possesses also a “vertical structure”. This is not only caused by the grammatical prescriptions – there are none for syllable types and some other units – but also by the type of the text, language, and some intuitively followed fixed structures which suggest the assumption that some law is working in the background. For rank-frequencies we found the well-known Zipf or zeta distribution, for some measured values we found the Hyperpoisson distribution, its special, limiting and modified cases but at the present stage there is no final answer. The modified cases indicate the existence of boundary conditions which exist for every law, but they are not known for the study of the “vertical behavior of texts” which is a quite new direction of research. We may put forward conjectures, set up hypotheses and test them. We may present the results to literary scientists and ask them how they would explain the deviations from the basic models. A more substantiated answer must wait until hundreds of texts are analyzed.

Nevertheless, we tried to show that text is also a vertically organized phenomenon and presented some ways of its examinations. Further studies would necessitate many texts and languages hence it is a task for future research. In any case, the results could be useful for typology, text sort analysis, the study of personal style, evolution of language and theoretical linguistics.

References

Altmann, G., Roelcke, Th. (2015). Morphological complexity of the word. *Glottotheory 6(1), 93-111.*

Hřebíček, L. (2000). *Variation in sequences.* Prague: Oriental Institute.

Kelih, E., Mačutek, J. (2013). Number of canonical syllable types: A continuous bivariate model. *Journal of Quantitative Linguistics 20(3), 241-251.*

Neumann, J.v., Kent, R.H., Bellinson, H.R., Hart, B.I. (1941). The mean square successive difference. *Annals of Mathematical Statistics 13, 152-163.*

Obradović, I., Obuljen, A., Vitas, D., Krstev, C., Radulović, V. (2010). Distribution of canonical syllable types in Serbian. In: Grzybek, P., Kelih, E., Mačutek, J. (eds.), *Text and Language: 145-157.* Wien: Praesens.

Popescu, I.-I., Kelih, E., Mačutek, J., Čech, R., Best. K-H., Altmann, G. (2010). *Vectors and Codes of Text.* Lüdenscheid: RAM-Verlag.

Wimmer, G., Altmann, G. (1999). *Thesaurus of univariate discrete probability distributions.* Essen: Stamm-Verlag.

Zörnig, P., Altmann, G. (1993). A model for the distribution of syllable types. In: Köhler, R., Rieger, B. (eds.), *Glottometrika 14: 190-196.* Trier: WVT.

Zörnig, P., Altmann, G. (2016). Consensus strings (in print).

Appendix

Slovak: E. Bachletová, *Moja Dolná zem.* Word length sentence-wise

[2,2,2,2,1,1,2,2,4,3,2,3,3,3,1,3,1,5,2,1,2,4,2,4]
[2,4,2,1,2,2,1,3,1,2,3,1,5,2,1,4,1,3]
[1,1,3]
[1,2,2,1,4,2]
[2,1,2,2,2,3,3,1,1,2,3,2,2,2,3,5,2]
[3,2,1,2,3]
[3,2,2,1,2,4,5,3,2,3,2,3,2,1,2,3,5,4,3,2,1,1,1,3,3,3,2,3,4,2,3,1,3,4,3,2,2,2,1,3,2,4,2]
[4]
[1,1,2,3,1,3,3,4,3,2,3,3,1,2,4,3,2,3,2,1,2]
[2,2,4,2,4,3,4,3,4,3,5,3,2,3,1,2,1,1]
[3,4,4,2,2,1,2,2]
[3]
[1,3,2,2,2,2,1,4,5]
[2,2,3,2]
[2,1,2,2,2,2,4,2,1,3,3,1,4]
[2,2,3,3,1,2,2,1,3,1,1,1,1,1,2]
[1,3,2,2,1,3,1,1,6,1,2,4]
[2,1,1,3,4,2,3,1,4,2,2,2]
[1,1,1,1,3,3,2,4,2,2,3,2]
[3,3,1,2]
[3,4,2]
[1,2,1,2,2,1,4]
[3,4,2,2,1,2,2,1,2,3]
[2,2]
[3,2,2,3,2,2,1,3,3]
[1,1,2,3,1,5,1,1]
[3,2,2,1,1,3]
[1,2,2,1,2,3,2,2,3,6,4]
[2,2,3,3,3,2,3,2,1,5]
[4,2,2,1,2,1,1,2]
[1,1,1,2,1,3,2,2]
[2,1,2,4,1,3,2,1,2,4,4,2,2,3]
[1,3,1,4,2,2,1,2,3,1,1,3,3,1,4]
[2,3,1,2,3]
[2,3]
[5]
[5,2,2,2,4,3,3,2,2,3,3]
[3,1,2,3,1,1,1,2,1,3,2,3,2,3,3,2,4,3]
[2,1,1,2,3,1,2,2,2]
[2,2]

[3,1,2,2,3,3,1,2,2]
[3,1,3,1,2,3,2,4,3,2,2]
[2,1,1,5,1,2,2,2,2,3,2,5,2,2,3]
[1,5,1,2,2]
[1,3,1,3,2,1,3,4]
[1,2,1,3,1,3,4]
[2,2,2,2,1,2,1,4,2,3,2,3,3,2,1,2,1,3,4,4,1,4,2]
[1,1,3]
[2,1,2,4,2,3,3,2,3]
[3,1,3,3,3,1,2,2,2,2,3,2]
[2,3,2,2,3]
[3,2]
[4,2,1,2,2,2,2,1,2,4,1]
[1,1,1,1,2,1,1,2,1,3,2,1,2,4,1,2]
[1,2,2,1,3,4,2,1,1,3,1,2]
[3,1,1,1,4,3,1,3,2,3,3,2]
[3,1,2,2,1,2,2]
[3,2,4,3,3]
[2,1,4,2,1,3,4,1,2,2,4]
[3,1,1,2,1,4,5,2,2,2,2,3,2,5,2,3,2,2,4,1]
[1,1,3,2,4,1,2,1,3,5,3]
[2,1,3,3,3,1,4,2,1,4,3]
[2,2,3,3,3,2,2,3,4,2,4]
[4]
[1,2,3,4,1,5]
[4,1,3,3,3,2,3,1,3,2,4,1,1,2]
[4,2,2,2,4,1,2,2,1,3,3,2]
[4,1,3,2,2,2,1,1,3,1,3]
[3,1,2,3,2,1,2,1,4,3,5,2,1,4,2,1,3,1,2,3,2,2,1]
[1,2,1,1,1,2]
[5,3,1,3,2]
[1,1,3,4,6,3,2,1,1,2,2,3,4,2]
[1,3,4]
[2,1,2]
[1,4,2,2,3,2]
[2,1,2,2,2,1,2,4,4,2,3,2]
[1,3,4,3,2,3,2,2]
[4,3]
[3,2,2]
[1,1,2,2,2]
[1,1,1,3,1,2,4,2,1]
[2,1,1,4,3,3,3,3,3,4,1,1,2,2,1,1,2,1,2,1,2,2,2,1,3]
[2,1,4,2,3]
[1,1,2,2,3]

[1,2,3,3,2,1,2,3]
[2,1,2,1,2]
[2,1,1,1,2,2,4,3,3,3,1,1,2,4]
[1,2,3]
[2,2,5,1,2,3,2,1,3]
[2,4,2,2,1,3,1]
[1,2,2]
[3,1,2,2];

Parts of speech in Chinese journalistic texts

T1: Multiple images of local officials - a hot issue in the political science in recent years (From Beijing Daily April 20, 2015)

T2: Can Internet make the wedding consumption more transparent? (From Consumer Daily April 15, 2015)

T3: Those people who illegally built shantytowns in the air become powerless - "limit down" verdict will be issued today, and demolitions will be started in mid May (From Beijing Daily April 17, 2015)

T4: The reforms of the burial customs in Hainan: the comfort for the living people and the dignity for the dead (From China Society April 14, 2015)

T5: The Central Bank of China has decided to drop the deposit reserve of residents for 1 percentage point, with the aim to stabilize the economic growth (From Beijing Daily April 20, 2015)

T 1 Chinese

[Av,N,V,N,V,U,Num,N,N,N,N,U,B,N]
[N,N,V,N,N,V,C,V,V,N,V,N,N]
[N,N,V,N,Adj,R,Q,N,T,Av,P,V,U,V,N,P,V,N,N,N,V,C,V,N,f,U,N,N]
[T,N,N,Av,Av,V,N,V,U,N,R,N,V,U,N]
[P,R,N,V,Adj,C,N,Av,B,U,N,N,N,U,N,V,B,Av,V,Av,V,U]
[P,N,U,S,R,V,V,N,N,U,Adj,N,Av,V,V,f,V,N,f,V,N,U,V,N,Av,V,V,N,U,Adj,N]
[P,S,N,N,S,R,C,V,R,S,V,U,V,V,N,U,N,N,N,Av,V,V,P,N,P,R,N,N,V,N,U,N]
[C,P,R,U,V,f,N,N,V,V,N,Adj,V,Adj,N,Av,Adj,U,N,Av,V,V,N,N,Adj,Num,V,Num, Adj,V,Adj,U,V,V,Av,V,V,N,N,N,U,N,Av,V,V,V,Adj,N,Adj,Adj,P,N,U,N,N]
[N,N,U,B,N,Av,V,z,U,N,N,U,Adj,N,P,Adj,N,N,N,P,N,U,N,N,P,N,P,V,P,N,N,U,N,f,V,N,N,Adj,U,N,N]
[N,N,N,U,B,N,V,Num,Q,N,V,U,R,U,Num,Q,N,Av,V,V,Num,Q,N,V,N,N,f,U,V]
[T,U,N,N,N,C,Av,V,Num,,N,Adj,N,V,f,P,V,V,V,U,N,N,Av,Av,V,Num,N,f,V,V,U,N,N,C,N,Av,Adj,N,N,Adj,U,Num,Q,Num,Q,N]
[N,Adj,V,P,V,V,N,N,N,N,N,U,Num,Q,N]
[N,N,Adj,V,P,N,N,C,N,N,V,U,N,V,V,P,N,N,N,N,N,N,C,N,U,N,N,V,N,N,N,N,C,P,B,N,U,V]

[P,N,U,f,V,N,V,R,P,V,P,N,U,N,N,N,V,V,N,f,Av,V,U,R,Av,Adj,U,N]
[V,Adj,C,Adj,V,P,V,Adj,C,Adj,V,V,U,C,V,V,U,B,C,B,P,N,V,V,U,C,V,C,V,U,C
,Adj,N,V,N,N,N,U]
[Av,P,R,Num,Q,V,N,f,V,Av,Adj,U,V,N]
[Num,N,C,V,U,Adj,V,B,N,V,V,V,Adj,U,N,f,C,P,Adj,U,N,N,C,Adj,N,f,N,U,V,V
,V,U,N,Av,V,V,N,U,V]
[C,P,R,U,V,U,N,N,V,V,U,N,R,N,Av,V,V,Adj,N,U,B,N,P,V,N,f,U,N,V]
[Num,Q,N,V,N,N,N,Num,Q,N,U,V]
[P,N,N,N,V,f,R,Av,V,Num,Q,Adj,N]
[N,U,Adj,N,V,U,N,N,V,N,R,f,N,P,V,N,,N,C,N,N,N,U,N]
[R,V,N,N,U,V,Av,V,P,N,C,V,P,N,C,N,N,Av,U,V,R,N,N,N,V,V,R,N]
[C,R,V,V,V,Adj,U,V,Av,V,N,K,V,f,B,K,V,f,C,f,V,V,Adj]
[R,V,N,Av,Av,V,P,P,T,N,N,N,U,V,C,P,Adj,U,N,f,R,Num,Q,N,V,U,Av,Adj,Y]
[N,V,P,Num,Q,N]
[C,N,U,N,V,P,Num,N,V,V,U,V,U,Adj,U,N,N,N]
[C,N,V,Num,N,U,C,P,N,N,U,V,V,U,V,f,V,Num,Q,U,N,N,N,U,N,N,N,V,U,V,
V,U,Adj,U,N,N]
[Num,N,N,U,V,C,N,V]
[P,V,V,N,V,C,N,V,U,V,N,N,f,N,Adj,V,N,N,V,N,N,Av,V,Av,Adj,U,V,N,N,Av,V
,U,Adj,U,N,P,f,N,V,U,Num,N,Adj,N,U,N]
[N,N,C,N,N,U,N,V,Adj,U,N]
[N,K,P,V,N,N,N,U,N,f,Av,V,V,N,N,U,B,N,C,V,P,R,V,Num,N,N,N,C,N,U,V,N,
Av,P,V,R,N,V,U,V,N,N,C,V,N,N,N,N,V,K,P,V,K,N,N,N,V,K,U,V,V,V,N,N,C,
N,N,U,N]
[R,N,Av,V,V,U,N,N,,N,R,Num,Q,N,V,C,V,U,V,V,P,Num,Q]
[Num,Q,P,N,N,V,N,N,U,N,C,N,N,U,N]
[Num,Q,P,N,N,V,N,N,U,N,K,C,V,K]
[Num,Q,P,N,N,V,f,N,U,V,N,U,V,P,S,N,N,V,V,V,C,V,U,N]
[V,R,Num,Q,N,Av,V,V,Num,Q,V,N,V]
[N,N,U,V,K,C,V,K]
[N,N,U,V,K]
[f,V,C,P,f,N,V,U,N]
[C,P,N,V,N,V,N,U,U,V,N,N,V,V,Num,Q,N,N,C,R,V,U,Adj,N,V,Adj,U]
[V,N,N,P,T,N,V,f,U,Adj,N]
[R,N,C,V,U,N,N,N,N,N,N,P,Adj,N,f,U,Adj,N,C,Adj,V,U,N,V,C,N,V,U,N]

T 2 Chinese

[N,V,V,N,V,V,Av,Adj]
[R,N,N]
[Av,V,Adj,U,N,Av,V,V,N,U,B,V]
[N,V,N,Adj,Num,Q,f,Av,V,N,Av,Av,V,N,U,N]
[P,N,Adj,P,N,N,V,U,Adj,N,f,P,N,Av,V,N,V,N,Av,V,U,R,Q,N]

[C,N,V,N,U,N,N,V,P,V,V,C,V,V,V,N,V,V,V,Av,Adj,U,N,Y]
[N,V,Av,Adj,T,N,z,U,Num,N,Av,V,V,N,U,N,N,T,Av,V,N,K,P,V,N,V,U,Adj,U,
N]
[Adj,N,V,N,N,Av,V,V,N,T,V,V,N,Av,Num,Num,Q,Y,R,Av,V,N,Av,Adj,C,Av,
V,N,Av,Av,V,V,R,Adj]
[R,V]
[V,V,B,N,f,U,Num,Q,Av,Adj,V,V,Av,V,U]
[N,C,Av,V,V,V,Av,U,Adj]
[B,N,C,N,U,N,Av,Av,V,R]
[Av,V,V,V,N,N,N,V,R,f,N,Av,Adj,Av,Adj,C,V,V,Q,N,R,V,N,V,N,V,V,N,Av,V,
Av,V,U,V,V,N,Adj,U,N,P,R,R,U,Adj,N,Av,V,Av,V]
[R,U,N,Av,V,V,V,U,N,V,N,V]
[R,U,N,C,N,Num,C,V,V,Av,Adj]
[C,T,P,V,N,U,N,Av,V,N,N,V,R,Num,N,V,V,N,V,C,Av,Adj]
[P,N,N,P,N,N,V,Num,Q,R,V,Av,Adj,U,V,N,N,P,N,S,V,Adj,N,N,Num,Q,N,Av,V
,Num,Q]
[N,N,Av,V,V,V,P,R,V,V,U,N,Av,V,N,V,N,C,N,V,N,N,N,V,N,V,V,P,N,f,V,Y,V,
V,Av,U,Num,Q,N,Av,V,V,N,V]
[Av,V,P,V,N,V,Num,Q,N,Av,Av,V,Num,Q,C,N,N,P,R,V,Num,Q]
[V,N,V,N,V,R,N,Av,P,R,V,f,N,Av,V,Adj,V,V,V,U,V,V,N,Av,Adj]
[R,Av,U,V]
[P,N,N,V,N,V,N,N,V,V,Num,C,V,N,N,Y]
[P,N,N,V,Num,N,V,Adj,P,N,N,N,V,V,N,V,N]
[C,N,V,U,N,N,Num,Q,V,Q,V,N,Av,V,U,N,C,Av,Av,V,P,V,U,N,f,V,Adj,C,V,U,
N,V,N,V,U,Adj,Adj,N]
[N,N,N,N,N,P,V,N,V,Num,Adj,N,N,V,N,Av,V,V,V,f,V,N,N,C,Av,V,R,V,N]
[N,V,V,N,V,N,P,N,V,N,U,Av,V]
[P,N,N,V,N,Adj,R,P,N,V,N,P,Num,Q,C,Num,N,V,Num,Q,V,N,Av,Num,N,f,N,R
,V,V,P,Num,Num,Q,f]
[R,Adj,U,N,Av,V,V,R,N,V]
[V,P,N,Av,Adj,N,Av,Adj,V,N,U,N,N,P,N,U,N,V,N,Av,Adj,U,N,V,V,N,N,U,N]
[V,Num,N,f,N,N,V,N,U,Adj,T,U,N,V,V,U,Adj]
[Av,V,N,U,V]
[Num,T,N,N,V,N,V,Num,N,Num,N,V,V,V,N,N,C,N,N,V,V,Num,P,Num,N,Num
,T,Av,V,R,P,N,N,V,N,N,U,V,K,V,Num,R,V,N,Adj,R,V,V,Num,Num,U,Adj,N]
[P,U,f,V,U,N,N,V,N,T,Num,N,Av,V,N,V,N,V,N,V,C,N,V,N]
[N,V]
[N,N,Av,Av,V,N,T,N,N,N,V,N,Num,Num,Q,Q,V]
[N,V,N,N,V,V,U,Num,Q,Q,V]
[V,N,N,V,V,N,V,Num,Q,U,N]
[V,V,V,Num,Av,Adj,U,N,R,N,Av,V,V,C,N,V,N,Av,V,N,N,U,V,V,C,V,Av,V,V,
R,N,N,N,V,Av,Adj,U,N]
[P,N,f,N,U,N,C,N,U,N,V,P,S,V,N,V,N,f,V,U,N,V,V,V,C,V,U,N,C,Av,V,N,K,Av
,Adj,Adj,U,V,C,V,U,Adj,Av,V,V,N,N,N,Num,V]

[R,P,V,N,U,Adj,V,U,Av,V,Adj,U]
[V,Num,N,N,R,P,V,N,N,N,V,Num,Num,Q,N,V,V,R,N,T,N,N,Av,V,V]
[R,Q,N,Av,V,V,C,N,V,N,Av,V,N,Adj,C,V,Av,Adj,V,V,N,Av,V,Num]
[R,C,V,N,Adj,Av,Adj,Av,V,N,N,V,N,Adj,Av,V,V]
[P,N,N,N,U,Num,V,U,N,V,N,V,V]
[V,Adj,U,Num,Q,V,V,P,Av,N,V,V,V,Adj,U,N,Av,V,P,Av,N,V,V,Adj,U,V,N,V,
Av,Adj]
[P,R,N,N,N,N,N,V,C,V,U,N,N,Adj,P,V,N,V,N,N,N,U,V,N,Av,V,U,N,V,N,N,Av
,Adj,U,N,Av,Adj,N,V,N,V,N,V,U,V,N,C,V,N,V,N,V,U,N,N,N,C,Num,Q,V,N]
[N,N,V,V,U,N,N,f,R,N,N,U,V,N,Adj,V,Av,Av,V,N,N,N,U,V,R,V,U,B,V]
[V,U,N,V,N,R,V,Q,B,N,U,N,N,Av,Adj,C,V,R,N,U,N,N,Av,V,V,Adj,U,B,N,N,V
,U,N,Adj,V,Av,V,V,N,U,R,N]
[C,P,N,N,Num,f,Num,f,Av,V,V,N,S,V,N,P,V,N,U,V,U,N,N,V,Av,Adj]
[C,C,V,N,Adj,Av,V,V,V,Av,B,U,V,C,Av,V,N,V,Av,V,N,U,N,N]
[T,Num,N,V,N,Av,V,V,N,V,N,U,V,N]
[P,R,N,N,P,N,V,N,U,V,K,Av,Adj,N,V,N]
[P,N,f,N,V,N,Av,N,Av,V,V,N,Adj,V,N,P,N,V,N,C,Num,Q,Adj,U,N,N,Av,V,Av,
Adj]
[B,V,N,N,V,V,Av,V,V,Av,Av,V,Av,V,V,V,Av,V,U,V,V,P,R,N,V,V,N,U,N]
[R,Num,V,N,V,N,N,V,U,N,V,N,V]
[P,Adj,N,V,N,V,N,U,V,V,V,V,U,Av,Adj]
[P,V,N,N,N,V,P,N,N,N,Num,N,C,R,N,V,N,V,P,N,N,C,N,N,f,Av,V,N,N,N,U,B,
N,Av,V,V,N,C,V,N,V]
[P,N,U,V,Num,N,Av,V,B,U,V,C,V,N,Av,Av,P,V,Adj]
[C,T,V,Av,V,Num,Q,N]
[P,V,f,U,f,Num,Q,V,N,U,V,Av,V,P,V,N,N,N,Av,Av,Adj,V,Num,Q,V,Num,Num
,Q,P,T,U,N,U,V,V,N,P,N,f,V,R,N,U,N,Av,Av,Av,Adj]
[T,N,Av,Av,V,V,U,N,V,N,f,V,R,Av,P,R,Av,Adj,U,N,V,P,V,N,U,V,P,N,f,V,U,V
,N,V,Av,Adj,U,V,Av,P,T,f,V,V,V,N,Av,V,V]
[N,N,V,N,Av,V,N,T,N,f,N,N,Adj,V,V,N,V,N,V,U,N,N,Av,Adj]
[N,f,Num,N,V,N,V,V,V,P,B,U,N,N]
[Num,N,N,N,C,V,N,N,N,V,Adj,N]
[C,T,Av,V,T,Av,V,P,V,U,N,N,P,N,Av,Adj]
[C,N,N,N,Adj,U,Av,Av,V,P,R,U,N,V,R,V,U,N]
[C,N,N,N,Adj,U,N,f,R,V,V,R,U,N,P,R,V,V,N,Av,Adj,U,N,V,V,N,U,Adj]
[R,N,P,R,N,f,Av,V,V,N,U,N,N]
[N,N,V,V]
[Adj,V,N,N,N,U,N,T,P,Adj,f,Av,V,V,U,V,U,N]
[P,Adj,U,V,N,N,N,V,N,N,V,V,U,V,U,Adj,U,N,N,C,V,V,V,Num,N,N,N,Av,V,V,
Num,Q,V,U,N,N,V,U,B,U,Adj,V,Adj,U,
N]
[N,U,N,Av,V]
[N,f,N,U,N,P,N,V,B,N]
[T,V,V,N,N,V,N,Av,V,V]

[R,Adj,U,N,V,Num,Q,T,V,N]
[N,U,N,V,V,V,V,U]
[V,N,f,U,N,V]
[C,R,V,U,N,Adj,U,N,C,R,Av,Av,V,U,N,V,V,N,N,P,R,N,f,R,V,V,Num,U,N,V,V,
Adj,U,N,P,V,R,U,N,P,R,Num,N,U,V,C,Av,V,Num,z,U,N,N,V,V,R,Adj,U,N]

T 3 Chinese

[T,V,Av,V,V,V,N,Num,N,T,Av,V,Av,V,S,V,V,N,U,N,K,V,U,R,N,N,Num,N,
Num,N,R,P,N,V,S,V,V,N,V,V,Num,N,V,N,N,N,Num,Q,N,N,N,N,V,Adj,N,V,N,
f,N,V,N,P,R,U,V,N,Av,V,V]
[C,N,V,V,V,Num,Q,V,V,N,P,R,N,V,R,N,f,Av,V,V,Adj,U,N]
[T,N,V,N,Av,V,V,V,N,V,V,V,N,Num,N,f,V,V,N]
[C,R,V,V,f,V,V,K,Av,V,V,C,P,T,Num,N,R,V,Av,V,Num,Q,V,V,N]
[C,V,V,K,V,Adj,V,U,f,N,Num,Av,Av,V,T,Num,N,,N,N,U,V,N,Av,z,U,V,P,Adj,
N,V,V,K,Av,V,V,V,N]
[C,V,U,N,V,N,Av,V,N,Num,N,T,f,P,S,V,V,N,U,V,V,Av,V,V,N,N,Num,Q,N,N,
N,N,C,N,U,V,N,Av,Av,P,V]
[Av,V,Num,Q,V,V,N,Av,V,N,N,N,Num,N,f,Av,V,N,U,N,N]
[R,Q,N,U,S,N,Av,V,N,V,N,P,N,N,Av,V,Num,Q,N,U,N]
[T,N,V,V,U,T,Num,Q,N,f,V,N,f,U,Adj,N,P,N,P,N,Num,V,V,V,U,N,Av,V,P,S,B,
N,V,Adj]
[Av,V,P,R,V,Num,N,P,Num,Q,N,f,N,N,S,f,V,V,Num,V,U,N,R,V,C,R,V,Q,Adj,
N,U,Av,V,V,V,U,Av,V,V,N,Adj,V,N,V,V,U,N]
[Num,N,Num,N,T,Num,N,Num,Q,P,N,N,Num,Q,N,N,N,N,V,U,Adj,N,Av,P,N,N
,N,V]
[V,N,N,N,N,N,N,N,N,V,V,N,N,N,N,N,N,N,N,Num,Q,N,N,N,N,N,U,Num,Q,N,U,
V,N,Av,V]
[P,R,Q,N,f,R,V,U,Num,Q,N,Av,Adj,U,Num,V,V,V,V,N]
[T,N,V,N,N,N,Num,V,N,U,N,P,T,Num,N,V,Av,Av,V,V,R,V,V,V,N,V,V,V]
[P,V,N,Av,Adj,Num,Num,f,V,N,Av,V,V,U,N,N,P,R,V,V,N,P,S,V,U,N]
[S,N,V,P,P,N,N,N,U,V,V,N,f,V,U,Num,Q,N,V,C,V,U,N,N,Num,Q,N,N,Av,P,V,
U,Num,Num,Q,N,f,U,V,V,N,Av,V,Num,Q,V,N,Num,N,Av,V,Num,Q]
[P,N,V,N,V,R,V,N,Av,Num,V]
[V,N,V,N,V]
[V,N,V,V,T,Av,V,Num,Q,V,V,U,N,Q,V,V,N,Av,V,V]
[V,N,N,Num,Q,N,S,N,Num,f,V,N,U,V,V,N,N,Av,V,V]
[V,N,V,N,N,Av,V,P,R,N,V,V,N,N]
[C,N,Av,V,V,U,N,P,V,V,P,T,R,V,Av,P,N,N,Av,V,V,Av,V,V]
[C,P,V,N,V,R,N,U,N,R,V,Av,Adj,U,V,f,N,U,N,V,V,N,N,V]
[V,Num,V,V,K,V,V,U,P,N,N,V,V,N,U,V]
[V,V,N,P,V,U,N,N,N,V,R,V,P,Num,Q,N,S,Av,V,N,Av,V,U,z,V,U,N,N,Num,V,
U,N,V,V,U,N,Av,V,P,V,Q,R]

[R,V,N,S,U,N,N,Av,V,Av,Adj,P,f,V,V,N,U,N,R,Av,V,V,V,V,N,Av,Adj,U,V,N,
R]
[N,U,V,U,N,Av,V,V,V,N,U,Num]
[Num,Num,N,f,R,V,Av,V,U,N,P,V,V,N,B,N,N,V]
[N,Av,V,N,N,V,N,V,N,P,Num,N,Num,N,Num,N,V,V,U,V,V,N,V,N,T,Av,V,P,N
,N,N,f,V,Num,Q,N]
[N,P,N,V,U,N,f,V,V,T,B,V,P,N,N,U,V,N,N,C,N,P,V,P,N,V,V]
[C,V,N,N,U,U,N,N,Av,Av,V,P,R,V]
[Num,Q,V,K,P,N,V,U,N,V,R,N,V,P,N,Adj,V,P,N,C,P,R,V,N,U,V,N,V,V,U]
[T,Av,P,R,N,V,V,V,U,N,V,T,R,Num,Q,N,N,U,N,R,Av,Adj,Y]
[f,Y,V,V]
[R,U,N,R,V,V,V,P,R,V,V,N,f,N,N,U,N,N,Av,V]
[V,U,V,N,N,T,U,V,K,V,N,V,R,V,N,U,N,P,V,N,V,N,f,Av,V]
[R,N,N,V,P,V,U,N,U,Adj]
[P,N,N,V,N,R,P,R,V,U,Num,Num,Q,S,N,V,N,V,N,V,R,R,V,Num,N,U,N,V,U,N
um]
[C,Av,V,N,V,N,Av,V,V,V,N,U,N,Av,V,P,V,P,R]
[C,C,V,V,S,U,Num,Q,N,Num,V,U,Num,Num,Num,Q,N,Av,V,R,V]
[N,V,N,K,Av,V,V,N,P,S,Num,Q,Q,V,f,V,V,N,N,Num,Q,N,N,N,P,f,P,f,Av,V,N,
N,Num,Q,N,N,N,N,N,N,N,C,N,N,R,P,N,N,N,Num,Q,N,Av,V,V,N,R,R,Av,V,V,
U,V,N,V,N,Av,Adj,V,V,U,Adj,N]
[N,C,N,S,Av,V,V,V,N,N,N,C,N,U,N,N,N,f,Av,V,z,V,V,Num,Q,V,N,N,Adj,Adj,
N,N,U,N,f,Av,V,U,N,V,C,V,Adj,N]
[V,V,V,R,f,V,N,K,U,B,N,R]
[C,T,R,C,N,N,V,Adj,U,N]
[V,U,V,N,Av,V,V,U,R,Adj,U,N,N]
[C,C,Av,V,f,Adj,N,U,N,Av,Av,V,V,Y]
[Num,Q,V,N,V]
[P,T,V,U,V,Av,V,Num,N,V,V,N,V,V,V,N]
[R,V,N,Av,Adj,U,V,B,N,N,U,V,N,N]
[T,R,Av,P,N,N,N,V,V,U,Num,Q,N,N,C,Num,Q,Av,V,Num,Q,N,U,N]
[C,R,Num,V,N,V,V,V,U,N,Av,V,V,V]
[N,V,V,V,P,V,V,C,V,V,U,Num,Q,V,N,K,f,V,Num,Num,Q,N,Av,V,R,V,Av,V,U,
N]
[R,N,f,U,N,R,Av,V,U,N,V,R,N,V,N,V,U,N,U,N,V,V,Q,N,C,N,R,N,S,V,N,N,N,
V,V,R,N,Av,P,N,Adj,V,Av,V,Av,Adj,U,N,P,V,V,V,Av,V,Av,V,N]
[Adj,Adj,N,U,Adj,N,Av,Av,V,V,Q,Q,V,Adj,U,N]
[P,R,R,V,P,N,V,U,N,f,V,P,B,N,V,V,R,Adj,N,R,V]
[Num,Q,V,U,Num,N,V,N,U,N,V]
[N,V,N,V,V,V,V,V,P,N,V,U,N,V,P,T,N,V]
[Av,V,N,Av,V,V]
[R,V]
[V,N,f,R,N,Av,Av,V,P,U,Num,Q,B,N,N,S]
[P,N,V,N,f,V,V,N,P,V,V,V,U,N,V,V,N,V,V,R]

[C,R,Q,V,V,V,N,N,V,Av,V,N,V,V,U,Av,V,V,V,R,V,N]
[R,Q,N,V]

T 4 Chinese

[N,N,V]
[V,K,V,V,V,K,V,N,B,N,N,N,N,N,Av,V,V,K,V,Adj]
[R,Av,V,N,V,P,N,U,V,C,V]
[T,N,N,N,N,V,U,R,B,N,V,V,V,U,N,P,Num,N,f,Av,V,U,N,V]
[R,V,N,N,N,P,N,V,Av,V,N,N,Av,V,N,V,Num,Q,V]
[C,Num,N,Adj,N,U,Av,V,N,N,V,V,U,R,N]
[N,Num,V,V]
[Adj,V,U,V,N,N,N,V,V,U,N,N,N,Av,V,Av,V,V,Adj,V,Adj,N,N,N,V,N,N,Av,V,
N,C,N]
[V,V,N,N,V,N,N,N,N,U,N,V,N]
[C,P,R,Adj,f,V,N,U,N,Av,V,R,U,N]
[S,N,N,N,f,N,V]
[N,N,N,V,N,N]
[N,N,N,Adj,N,V,Adj,N]
[N,N,N,f,Adj,Adj,U,N,V,N]
[Num,N,Num,N,N,N,V,Adj,V,f,V,N,Num,P,Num,N,V,P,N,C,N,N,f,N,N,N,N,P,
N,P,N]
[N,N,N,N,N,N,V,N,N,N,V,V,N,f,U,N,V,Num,Num,Q,Q,V,Av,Num,Num,Q]
[Av,V,Av,V,U,N,C,V,Num,N,Av,V,U,N,U,Adj,Adj,V,N,N,V,N,V]
[Num,N,V,N,N,N,V,U,P,N,V,U,V,C,P,N,V,V,Num,N,V,U,N,V,Av,V,Av,V,V,N,
V,P,Adj,V,P,Adj,V,P,Av,V,Av,V,P,V,V,P,N,V,N,P,N,V,N,U,V,
V,N,N,N,N,U,N,V,Adj,N,Adj,V,V,U,N,N,V,U,N]
[P,Adj,V,Num,N,V,N,Av,V,Av,V,N,Av,V,N,V,Av,V,V]
[Adj,N,V,V,N,Num,P,Num,N,V,V,N,f,Av,V,Adj,N]
[P,Av,V,V,N,N,V]
[N,N,V,V,N,N,Av,V,P,V,V,N,N,V,U,N,N,N,N,V,N,C,N,N,V,N,U,N,V,N,V,V,N,
V]
[P,V,Num,N,Num,T,Adj,N,V,N,N,Num,Q]
[Av,V,N,N,V,Num,Q,N,N]
[V,Av,V,Av,V,N,Num,Num,Num,Q,V,N,N,N,Av,Adj]
[V,N,V,Adj,N,N,V,V,K,Adj,V,P,N,f,R,N,U,N,N,Av,V,Adj,U,R,Q,N,Av,V,R,V,
N,N]
[N,V,Num,Adj,V,U,N,V,U,V,N,U,N,C,Av,V,V,K,U,N,C,V,N,Av,V,N]
[T,U,N,N,Av,V,V,N,Adj,N,V,f,Av,V,V,U,V,N]
[N,V,P,V,Adj,N,P,N,N,V,N,C,V,f,N,Adj,P,R,V,V,N,N,Num,N,P,N,N,N,U,N,N,
V,U,V,N,N,N,V,V,P,N,N,V,f,V,N,V,V,N]
[Num,T,V,N,N,N,V,Adj,N,V,U,N,B,N,P,N,U,V,N,R,Num,N,P,V,N,f,V,V,N,N,V
,Adj,U,Num,Adj,N]

[V,V,Av,V,N,Av,V,Adj,U]
[N,N,B,N,N,V,N,Av,V,Av,V,N,Num,Num,Q,Adj,V,N,N,N]
[R,P,R,V,V,Av,V,Av,V,U,V,N,V,N,N,f,V,N,R,U,N,N,Adj,V,V,V,V,N]
[Adj,N,Av,V,N,V,N,C,Av,V,Adj,N,Adj,V,C,V,N,Av,Adj]
[T,N,N,V,U,V,N,V,N,N,V,Av,V,C,V,Av,V,N,V,N,V,V,f,f,V,V,N,C,R,V,N,V,N,
U,N,Num,P,N,N,Adj,N,Av,V,N,V,N,U,R,Q,N,V,Num,Q]
[N,Av,Av,V,V,N,N,N,V,U,N,V,N,N,N,B,N,P,N,P,N,U,N,Adj,N,N,N,V,N,V,N,C
,P,N,N,V,f,U,B,N,P,N,P,N,V,N,V,U,V,Num,Q,N,V,V,N,V,V,P,N]
[V,Num,T,N,R,N,N,V,S,V,N,Av,V,U,V,N,N,N,Av,V,P,N,V,V,C,V,P,V,N,N,U,
V,N,V,V,U,N]
[P,V,Av,Num,N,R,N,N,Av,V,N,V,N,N,N,N,P,Num,Num,Num,Q]
[N,N,V,N,N,V,N,N,V,N,V,R,U,V,V,N,P,N]
[N,U,N,N,Av,V,Av,V]
[V,V,C,V,P,N,U,N,V,V,N,U,R,Num,Q,V]
[P,N,N,U,V,N,U,V, N,P,N,V,U,Av,Adj,Av,Adj,U,N]
[P,N,V,Num,N,U,N,N,C,V,U,N,P,N,U,N,Av,V,U,N,P,N,U,V,V,V,Num,N,U,N,
N,V,N]
[C,N,N,Av,V,Adj,N,N,U,V,C,V,V,N,N,P,V,N,V,V,N,N,Adj,V,V,V,V,V,N,P,N,
N,V,N,N,N,V,U,V,
P,N,V,N]
[T,T,N,N,N,N,N,Av,V,V,P,Adj,V,N,V,U,V,N,V,N,N,V,V,Adj,N,V,V,V,Adj,V,P
,V,N,N,V]
[N,N,V,Av,Adj,C,V,Num,N,V,Av,V,V,N,V,N]
[P,N,N,V,Av,V,V,N,N,N,N,N,V,V,Adj,V,N,N,P,N,N,V,P,N]
[N,N,N,N,N,B,N,N,V,P,N,N,V,N,N,N,B,N,N,N,N,N,B,N,N,U,N,N,Av,P,N,V,N]
[P,R,U,V,f,N,N,N,Av,V,V,N,Av,P,N,V,V,Adj,N,N,Adj,N]
[N,V,V,N,U,N,V,V,N,B,N,U,N]
[P,S,N,N,Av,Adj,Adj,C,Adj,Adj]
[N,P,R,V,V,Av,V,U,N,V,N,V,V,C,V,K,Av,P,R,V,V,f,U,N,C,N]
[R,V,Num,Q,N]
[T,R,N,Av,P,N,Adj,V,V,N,U,N,N,V,Adj,U,V]
[N,N,N,N,V,N,N,Adj,V,N,Av,V,N,V,N,f,V,U,Num,N,N,Adj]
[R,Adj,V,V,Adj,N,V,N,V,Y,Av,V,Num,Q,N]
[N,V,B,N,V,N,N,N,f,Adj,Adj,Av,V,N]
[V,K,Adj,V,P,R,Av,V,V]
[N,N,P,V,f,Av,Av,V,N,Adj,V,U,Adj,N,N,Av,V,U,Adj,V,P,S,U,N,N,V,N,f,N,Av,
Av,Av,V,f,R,Av,V,N,f,V,N,Y] [N,V,N,V,N,N,V,N,N,Av,V,V]
[N,V,T,N,N,Av,Av,V,N,V,N,N,V,Av,V,U,N,V,N,Adj,V,N,V,N,V,N,V,N,V,U,N,
N,V,N,N,V,N,N,N,Av,V,V,N,Adj,U,N]

T 5 Chinese

[N,V,N,Num,Q,Num,V,V,C,Av,V,N,C,Av,Av,V,Num,Num,Num,Q,N,R,N,N,N,
T,T,N,V,P,Num,N,Num,N,Num,N,V,V,R,N,Q,N,N,N,N,K,Num,Q,Num]

[N,N,K,V,f,B,N,C,B,N,Av,Av,V,Num,C,Num,U,V,N,K]
[C,P,V,P,Adj,Adj,N,Num,N,C,Adj,N,V,U,U,V,N,Av,V,Num,N,Av,V,N,N]
[Av,V,R,V,N,Av,V,V,N,Num,Num,Num,Q,Num]
[N,V,V,N,R,V,N,Av,P,V,V,U,N,V,V,V,U,N]
[N,Adj,U,N,N,Av,V,U,Num,Q,N,Y]
[V,Num,Q,Adj,N,V,V,Av,V,V,V,Num,Q,Adj,V,V,T,V]
[P,V,V,V,V,N,N,V,T,V,N,U,N,N,V,U,Av,Adj]
[V,P,N,N,U,V,V,N,Num,Q,N,Av,V,Av,V,U]
[N,R,V,N,N,T,Av,V,V]
[C,Adj,V,V,Num,Q,Num,C,V,P,Num,Q,Av,V,N,Av,V,Av,V,U,N,V]
[T,N,Num,N,Num,N,U,V,N,Av,V,U,Num,Q,Num]
[Av,V,N,N,V]
[P,Num,N,Num,N,V,Av,Av,P,N,N,N,U,N,N,Av,V,N,N,N,K,Num,Q,Num,C,Adj,V,N,V,N,N,N,K,P,N,N]
[P,N,N,V,N,Av,V,N,N,N,K,Num,Q,Num]
[P,V,Adj,V,N,C,Num,N,C,Adj,Adj,N,N,V,B,N,U,V,N,C,N,N,N,V,V,Av,B,N,B,N,V,Num,Q,Num,U,N,N,K]
[N,N,V,V,N,P,V,N,V,N,C,N,V,N,C,V,U,V,P,N,U,N,N,V,U,N,N,V,R,N,N,U,N,Av,V,N,N,K]
[V,N,V,U,N,R,P,N,N,V,Num,Q,N,C,N,N,V,Adj,V,P,N,Num,P,N,N,C,Av,Adj,V,U,Num,Num,N,V,V,V,P,N]
[V,N,U,N,V,N,V,V,P,N,N,Av,V,Num,N,C,Adj,Adj,N,U,V,N,C,V,V,V,P,V,V,R,N,V,N,U,Av,B,N]
[N,V,Num,N,N,N,Av,V,V,Num,V,V,C,V,Num,Q,N,Av,Av,Adj,Av,V,V,N,V,V,N,V,N,Av,Av,Adj]
[V,V,Num,Q,Num,C,Av,V,T,U,Num,Av,V,N,P,N,V,N,U,N,P,V]
[N,N,N,C,N,N,N,N,V]
[T,N,N,N,N,Num,Num,Num,Num,Q,V,R,V,Av,C,Av,V,N,Av,V,N,Av,Num,Num,Num,Q]
[N,N,N,V,N,N,N,N,V,R,Av,Av,V,N,N,V,N,N,V,V,N,V,N,V,C,V,B,N,V,V,C,V,N,V]
[Av,T,Num,N,Num,N,C,T,Num,N,Num,N,N,Num,Q,V,V,N,N,V,U,V,N,Av,V,N,V,N,Av,V,V]
[N,N,V,Num,N,Num,N,f,N,V,N,V,Num,P,T,f,V,Num,Q,N,P,T,f,V,Num,Q,N]
[N,Av,V,V,T,V,N,V,Av,Adj,T,V,V,R,T,V]
[T,V,V,N,Av,V,T,V,V,R,Adj,V,U,V,V]
[R,V,N,P,V,N,Y]
[V,N,N,Num,V,N,K,Av,V,N,C,N,Av,V,V,N,Y]
[P,T,T,V,V,U,N,N,N,V,V]
[N,N,V,V,R,V,N,Av,V,U,V,N,N,N,V,U,N,Av,V,V,V,Av,V,N,N,N,V,z,U,N,N,N,V,U,N,Av,V,V,N,V,Av,Adj,V,V,N,N,B,N,V,Adj,V,Num,N]
[P,V,N,U,Av,V,N,Adj,K,U,N,C,N,N]
[Num,N,Av,V,N,N,U,N,V,N,Av,V,V]
[C,N,N,Av,V,Av,V,V,Adj,V,N,U,N,N,R,V,N,Av,V,Adj]

[R,V,Av,V,U,T,N,P,Av,V,U,N,f,V,V,P,V,V,Num,N,V,V,Num,N,N,U,Adj,N,Av,V,V]
[P,f,T,U,N]
[T,R,V,V,Adj,V,Adj,V,T,R,V,V,Adj,V,Adj,V,T,R,V,V,Adj,V,Adj,V]
[N,V]
[f,T,N,V,N,N,U,Num,Q,V,N,V,V,U,P,V,V,N,N,V,V,N,U,N,C,P,N,N,V,V,V,N,N,V,U,Num,Q,N,V,V,P,R,N,V,N,N,V,N,N,N,U,V,U]
[V,R,V,V,U,N,U,N,N,V,Av,V,Av,V,Av,V,Adj,Y,V,N,V,V,V,N,Y,Av,N,Adj,U,V,V]
[f,T,T,N,N,N,N,V,V,V,V,N,N,Av,V,V,U,V,V,N,V,V,V,N,V,Adj,V,Av,Av,V,Av,V]
[C,N,U,V,U,Num,Q,N,V,P,V,N,U,V,C,V]
[Adj,V,Av,V,Adj,V,P,Num,Q,V,V,f,Num,Q,N,Adj,N,Av,V,V,V,Adj,N,Av,V]
[C,N,N,N,N,N,N,Av,V,B,N,N,R,C,Av,Av,V,Av,V]
[T,Num,N,V,N,f,U,Num,Q,N,B,N,V,Num,B,N,V,Num]
[C,Num,N,V,N,P,f,T,V,B,N,Av,V,Av,Num]
[C,V,N,V,N,V,N,N,V,N,V,Av,Av,Adj,V,Av,Adj,V,V,K]
[P,V,N,P,N,N,V,U,V,V,N,V,N,N,Num,N,V,V,V]
[R,V,V,V,N,V,U,N,V,V,N,V,Av,Av,V,V,K,V,N,V,P,f,Av,Adj,R,Av,V,N,V,N,N,Av,V,V,V,Num,N,Av,Av,Adj,V,N,N,U,N,Av,V,R,N,Av,Av,P,N,V,N,V,N,N,Av,V,V,Adj,V]
[N,V,V,N,N,N,V,V,R,N,V,N,C,T,N,Av,V,N,V,N,V,B,f,U,N,V,Num,Q,N,N,Av,V,B,V]
[N,Adj,Av,V,N,V,Av,Adj,Av,V,P,N,N,C,N,Adj,N,Adj,U,N,U,V,Num,Q,N,Av,V,V,V,U,N];

Parts-of-speech: German:
J.W.v.Goethe: Der Erlkönig, verswise

P,V,Av,Av,Pr,N,C,N,
P,V,Art,N,Pr,P,N,
P,V,Art,N,Av,Pr,Art,N,
P,V,P,Av,P,V,P,Av,

P,N,P,V,P,Av,Av,P,N,
V,N,P,Art,N,Part,
Art,N,Pr,N,C,N
P,N,P,V,Art,N

P,A,N,V,V,Pr,P,
Av,A,N,V,P,Pr,P,
P,A,N,V,Pr,Art,N,
P,N,V,P,A,N,

P,N,P,N,C,V,P,Part,
P,N,P,Av,V,
V,A,V,A,P,N,
Pr,A,N,V,Art,N,

V,A,N,P,C,P,V,
P,N,V,P,V,Av,
P,N,V,Art,A,N
C,V,C,V,C,V,P,Av,

P,N,P,N,C,V,P,Part,Av,
N,N,Pr,A,N,
P,N,P,N,P,V,P,Av,
P,V,Art,A,N,Av,A,

P,V,P,P,V,P,A,N,
C,V,P,Part,A,Av,V,P,N,
P,N,P,N,Av,V,P,P,Av,
N,V,P,Art,N,V,

Art,N,V,P,V,Av,
P,V,Pr,N,Art,A,N,
V,Art,N,Pr,N,C,N,
Pr,P,N,Art,N,V,A

Polish

ADJ adjecitve
ADV adverb
CONJ conjunction
N noun
PART particle
POST postposition
PREP preposition
V verb
VPRON verbal pronoun (*się*)
NA empty space (filler to reach 15 elements)

Polish POS text 1

[ADJ,N,PREP,N,PREP,N,PREP,N,PREP,N,N,ADJ,ADJ,N,V]
[V,ADV,N,N,CONJ,V,N,ADV,N,NA,NA,NA,NA,NA,NA]
[N,V,ADJ,N,ADJ,CONJ,N,ADJ,PREP,ADJ,PREP,ADJ,N,N,N]

[N,ADJ,CONJ,N,PREP,N,V,ADJ,N,NA,NA,NA,NA,NA,NA]
[PREP,N,ADV,V,PREP,N,PREP,N,ADV,ADJ,CONJ,ADJ,N,ADJ,ADJ]
[N,V,VPRON,CONJ,N,ADJ,N,ADJ,N,ADJ,N,ADJ,N,CONJ,V]
[ADJ,V,CONJ,N,ADJ,N,V,N,N,N,N,CONJ,N,PREP,N]
[ADJ,ADJ,N,ADJ,PREP,ADJ,N,V,ADV,PREP,N,ADV,ADJ,N,PREP]
[ADJ,N,V,VPRON,PREP,N,N,CONJ,V,ADJ,N,CONJ,V,N,ADJ]
[ADJ,N,V,N,PREP,N,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[PREP,N,N,N,ADV,V,N,NA,NA,NA,NA,NA,NA,NA,NA]
[V,CONJ,PART,N,N,PREP,N,N,N,N,N,NA,NA,NA,NA]
[ADV,V,VPRON,N,N,PREP,N,V,ADJ,N,NA,NA,NA,NA,NA]
[V,CONJ,ADJ,N,V,ADV,N,PREP,N,CONJ,N,N,N,NA,NA]
[PREP,N,V,CONJ,N,CONJ,ADV,ADV,V,VPRON,N,V,N,NA,NA]
[ADV,N,N,PREP,ADJ,N,PREP,N,V,N,N,ADV,N,CONJ,N]
[ADV,ADV,V,CONJ,ADV,V,VPRON,N,N,NA,NA,NA,NA,NA,NA]
[ADV,PART,ADJ,N,ADJ,ADJ,CONJ,ADJ,ADJ,N,CONJ,N,PREP,N,CONJ]
[ADJ,N,V,N,ADJ,N,ADJ,N,ADJ,ADV,N,CONJ,N,CONJ,ADJ]
[CONJ,V,PREP,ADJ,N,PART,V,ADV,PREP,ADJ,N,CONJ,N,ADJ,V]
[N,V,VPRON,N,ADV,PREP,N,N,N,ADV,ADV,V,PREP,N,V]
[ADV,PREP,N,N,ADJ,N,V,PREP,ADJ,N,CONJ,ADV,PREP,N,V]
[N,PREP,CONJ,V,PREP,N,V,N,N,PREP,N,ADJ,N,ADJ,PREP]
[N,ADJ,N,N,ADJ,PREP,N,PREP,N,V,VPRON,PREP,N,ADV,ADJ]
[ADJ,N,V,PREP,N,ADJ,N,NA,NA,NA,NA,NA,NA,NA,NA]
[ADJ,N,CONJ,V,ADV,ADJ,V,VPRON,PREP,N,N,ADV,ADV,ADJ,N]
[ADJ,N,PREP,N,V,ADV,N,PREP,ADJ,N,N,PREP,CONJ,V,ADJ]
[PREP,N,N,V,VPRON,PREP,N,ADJ,N,CONJ,V,ADJ,N,NA,NA]
[PREP,N,N,N,N,V,N,N,CONJ,N,N,V,PART,N,CONJ]
[N,PREP,N,PART,V,N,N,PREP,N,N,CONJ,N,N,N,NA]
[N,V,PREP,N,ADV,ADJ,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[CONJ,PREP,N,ADJ,N,N,ADJ,ADJ,V,N,N,PREP,N,ADJ,CONJ]
[N,N,ADJ,PREP,N,V,ADV,ADV,ADJ,NA,NA,NA,NA,NA,NA]
[ADJ,ADJ,N,ADJ,V,VPRON,PREP,ADJ,N,V,N,CONJ,V,PREP,N]
[ADJ,N,V,ADV,ADJ,CONJ,PREP,N,N,V,N,N,CONJ,ADV,PREP]
[ADJ,N,ADJ,V,N,PREP,N,N,ADV,N,N,N,NA,NA,NA]
[ADJ,ADJ,N,CONJ,N,V,N,N,N,N,ADJ,N,PREP,N,ADJ]
[ADJ,N,ADJ,N,ADJ,N,N,PREP,N,N,V,ADJ,PREP,N,PREP]
[ADV,PREP,N,ADV,N,V,N,ADJ,N,ADJ,ADV,V,N,N,N]
[ADV,V,ADJ,N,N,CONJ,PREP,ADJ,N,V,N,CONJ,V,VPRON,ADV]
[ADV,ADJ,N,PREP,N,V,ADV,N,PREP,N,CONJ,N,PREP,N,PART]
[ADJ,N,V,ADV,N,ADJ,CONJ,ADJ,NA,NA,NA,NA,NA,NA,NA]
[PREP,N,ADJ,ADJ,N,ADV,ADJ,N,N,CONJ,PREP,ADJ,N,V,VPRON]
[N,ADJ,N,V,ADJ,ADJ,ADJ,CONJ,ADJ,ADJ,NA,NA,NA,NA,NA]
[N,N,V,ADV,N,PREP,N,PREP,N,ADV,PREP,N,N,ADJ,N]
[ADV,VPRON,ADV,N,CONJ,N,V,ADJ,N,N,CONJ,N,ADJ,NA,NA]
[CONJ,V,N,VPRON,ADV,PREP,N,V,ADV,V,V,N,N,ADJ,N]

[CONJ,V,ADJ,N,ADV,N,ADJ,PREP,N,N,N,PREP,N,ADJ,NA]
[PREP,N,N,V,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[ADJ,PREP,N,PREP,N,N,ADJ,PREP,N,ADJ,N,V,ADJ,PREP,N]

Polish POS text 2
[N,PREP,ADJ,N,PREP,CONJ,N,V,N,V,ADJ,N,NA,NA,NA]
[PREP,ADJ,N,N,N,ADJ,ADJ,N,V,N,NA,NA,NA,NA,NA]
[PREP,ADJ,N,N,V,N,N,PREP,ADJ,N,ADJ,ADJ,CONJ,ADV,V]
[V,VPRON,PREP,N,N,CONJ,N,ADJ,PREP,N,ADJ,N,CONJ,N,ADJ]
[V,N,ADJ,N,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[ADJ,PREP,N,N,N,V,PREP,N,PREP,N,ADJ,N,NA,NA,NA]
[ADJ,N,PREP,PART,ADJ,N,ADJ,V,ADJ,N,N,NA,NA,NA,NA]
[N,N,N,PART,N,N,CONJ,ADV,NA,NA,NA,NA,NA,NA,NA]
[CONJ,ADJ,N,N,N,ADV,V,ADV,PREP,N,ADJ,N,ADV,PREP,ADJ]
[N,ADV,ADJ,N,PART,ADV,N,PREP,ADJ,N,ADJ,N,NA,NA,NA]
[N,V,ADV,N,N,ADJ,CONJ,V,ADJ,PREP,ADJ,ADJ,N,ADJ,N]
[ADJ,N,ADJ,N,N,V,N,N,CONJ,N,N,N,ADJ,PREP,N]
[N,V,N,N,ADJ,N,ADV,ADV,N,CONJ,N,ADJ,ADV,N,ADJ]
[PREP,N,ADJ,N,N,V,PREP,ADJ,N,N,CONJ,N,ADJ,N,NA]
[N,V,VPRON,ADV,PREP,N,ADJ,CONJ,V,PREP,N,N,ADJ,NA,NA]
[ADJ,PREP,N,N,CONJ,N,V,CONJ,PREP,N,CONJ,N,ADJ,PREP,N]
[ADV,PART,V,N,ADJ,N,ADJ,CONJ,V,N,ADV,N,PREP,ADJ,N]
[ADV,ADV,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[ADJ,N,CONJ,N,N,ADJ,V,CONJ,N,ADJ,CONJ,N,ADV,V,VPRON]
[N,V,CONJ,ADJ,CONJ,ADJ,N,V,N,ADV,ADV,CONJ,PREP,ADJ,N]
[ADJ,ADJ,N,V,PREP,ADJ,N,ADJ,ADV,PART,ADV,PREP,N,N,ADJ]
[ADJ,N,PREP,N,N,ADJ,V,CONJ,N,N,PREP,N,ADJ,N,NA]
[PREP,ADJ,N,PREP,ADJ,N,N,V,PREP,N,N,PREP,N,N,N]
[ADJ,N,PREP,N,N,ADJ,V,ADV,ADJ,N,ADV,N,ADJ,CONJ,N]
[N,V,N,ADJ,N,ADJ,V,ADV,N,N,CONJ,ADJ,V,N,ADJ]
[V,ADV,CONJ,ADJ,ADJ,ADJ,N,ADJ,CONJ,V,ADJ,N,ADJ,ADJ,V]
[N,ADJ,N,ADJ,PART,ADV,ADJ,N,N,NA,NA,NA,NA,NA,NA]
[PREP,N,N,N,PREP,ADJ,N,ADJ,N,V,N,N,ADJ,N,N]
[V,N,ADV,ADJ,CONJ,N,ADJ,V,VPRON,ADV,ADV,ADV,PREP,N,N]
[V,N,N,N,N,ADJ,CONJ,N,N,CONJ,PREP,ADJ,N,V,N]
[ADJ,N,N,V,VPRON,PREP,ADJ,ADJ,N,N,ADJ,PREP,ADJ,NA,NA]
[ADV,N,V,VPRON,N,ADJ,N,PREP,ADJ,CONJ,ADJ,ADJ,N,NA,NA]
[ADJ,N,CONJ,N,V,ADV,ADV,PREP,N,ADV,N,ADJ,V,N,CONJ]
[ADV,N,N,PREP,ADJ,N,CONJ,N,V,VPRON,PREP,N,CONJ,V,N]
[N,N,ADJ,N,N,V,N,N,PREP,CONJ,V,N,N,NA,NA]
[CONJ,PREP,N,ADJ,N,V,N,N,N,ADJ,N,ADJ,N,V,ADV]
[ADV,PREP,N,N,CONJ,V,VPRON,ADJ,N,CONJ,N,N,ADJ,ADV,ADJ]
[PREP,N,ADJ,N,N,ADV,N,V,N,ADJ,N,CONJ,V,ADJ,N]
[PREP,ADJ,N,V,ADJ,ADV,PREP,N,N,N,ADJ,PREP,ADJ,N,N]

[PREP,N,ADJ,N,ADJ,N,V,VPRON,N,ADJ,N,ADJ,CONJ,ADJ,V]
[PREP,N,ADV,ADV,ADJ,PREP,N,N,PREP,N,V,VPRON,ADV,ADJ,NA]
[PREP,N,ADJ,ADJ,N,V,ADJ,N,ADJ,ADJ,PREP,N,N,ADJ,NA]
[PREP,N,PREP,N,N,ADJ,PREP,N,PREP,N,ADJ,V,N,PREP,N]
[ADV,N,CONJ,PREP,ADJ,N,ADJ,N,V,N,NA,NA,NA,NA,NA]
[N,VPRON,V,ADV,PREP,ADJ,N,NA,NA,NA,NA,NA,NA,NA,NA]
[V,ADV,PREP,N,PREP,ADJ,N,ADV,N,N,N,CONJ,V,PREP,N]
[ADV,N,N,N,ADJ,PREP,ADJ,N,ADJ,PREP,N,ADJ,CONJ,ADJ,PREP]
[ADV,N,V,PREP,N,ADV,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[PART,V,ADV,N,ADJ,N,CONJ,V,N,N,ADJ,NA,NA,NA,NA]
[ADJ,N,N,ADJ,V,ADV,N,NA,NA,NA,NA,NA,NA,NA,NA]

Polish POS text 3
[V,N,VPRON,CONJ,N,ADJ,N,ADV,V,NA,NA,NA,NA,NA,NA]
[CONJ,PREP,N,PART,V,N,ADJ,ADJ,N,NA,NA,NA,NA,NA,NA]
[ADV,N,PREP,N,V,ADJ,CONJ,ADV,ADJ,N,NA,NA,NA,NA,NA]
[V,N,V,PREP,ADJ,N,ADV,N,V,N,CONJ,V,ADJ,ADJ,PREP]
[PART,V,CONJ,VPRON,V,PREP,N,N,CONJ,PREP,N,CONJ,V,CONJ,N]
[N,ADJ,N,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[PREP,ADJ,N,N,N,V,ADV,ADJ,N,N,CONJ,V,N,ADV,NA]
[ADV,ADV,V,CONJ,PREP,N,N,ADJ,N,V,ADJ,N,NA,NA,NA]
[V,PREP,N,ADJ,ADJ,N,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[N,N,PREP,N,N,PREP,N,ADJ,ADJ,ADV,N,NA,NA,NA,NA]
[V,ADJ,N,N,V,N,ADV,ADJ,N,N,ADV,ADJ,N,NA,NA]
[V,PREP,ADJ,N,PREP,ADJ,PREP,ADJ,CONJ,ADV,ADJ,NA,NA,NA,NA]
[ADJ,V,N,CONJ,N,V,ADJ,NA,NA,NA,NA,NA,NA,NA,NA]
[N,ADJ,N,V,ADJ,PREP,ADJ,N,ADJ,ADJ,N,PART,ADJ,N,CONJ]
[PREP,N,V,ADJ,N,N,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[V,N,ADV,ADJ,ADJ,N,CONJ,N,NA,NA,NA,NA,NA,NA,NA]
[PREP,N,N,V,ADJ,ADJ,N,ADV,PREP,N,N,PREP,N,PREP,ADJ]
[N,ADV,V,VPRON,PREP,ADJ,N,ADV,ADJ,N,ADV,ADJ,N,ADJ,CONJ]
[PREP,N,PREP,ADJ,N,CONJ,N,N,N,V,ADV,N,NA,NA,NA]
[PREP,N,ADJ,N,PREP,N,V,N,ADV,ADV,ADJ,N,N,ADJ,PREP]
[ADJ,N,V,VPRON,PREP,ADJ,N,ADV,ADJ,N,N,PREP,N,PREP,ADJ]
[ADJ,ADJ,N,CONJ,N,N,CONJ,N,ADJ,N,V,CONJ,N,V,ADV]
[ADJ,N,V,VPRON,ADV,ADJ,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[PREP,N,PREP,N,N,V,N,N,ADJ,N,ADV,NA,NA,NA,NA]
[CONJ,N,ADJ,N,N,V,PREP,N,ADJ,N,CONJ,ADV,V,N,N]
[PREP,ADJ,N,V,PREP,N,ADJ,N,CONJ,PREP,N,N,V,CONJ,N]
[PREP,N,N,V,CONJ,ADJ,N,PREP,ADJ,N,PART,V,ADJ,ADJ,N]
[N,N,PREP,N,N,V,ADV,PREP,N,ADJ,PREP,ADJ,N,NA,NA]
[N,N,PREP,N,CONJ,N,N,V,CONJ,N,ADJ,V,PREP,N,N]
[PREP,N,N,ADV,ADJ,N,V,ADJ,CONJ,ADV,ADJ,V,V,N,ADJ]
[V,VPRON,CONJ,ADJ,N,PREP,N,CONJ,N,ADJ,N,ADJ,V,VPRON,PREP]

[ADJ,ADV,V,PART,ADV,N,CONJ,ADV,ADJ,N,CONJ,ADJ,PART,V,N]
[ADJ,ADJ,N,V,ADV,ADJ,N,PREP,ADJ,N,CONJ,ADJ,N,PREP,ADJ]
[V,VPRON,ADV,N,ADV,PART,ADJ,N,V,PREP,N,N,PREP,ADJ,N]
[PART,N,ADJ,N,V,N,PREP,N,NA,NA,NA,NA,NA,NA,NA]
[ADV,N,VPRON,PREP,ADJ,N,ADJ,ADJ,N,ADV,ADJ,PREP,ADJ,N,NA]
[N,V,VPRON,CONJ,ADV,ADJ,ADJ,N,PART,ADJ,N,CONJ,V,ADV,N]
[PART,PART,V,N,ADJ,N,N,N,ADJ,N,N,NA,NA,NA,NA]
[ADJ,N,V,PREP,ADJ,N,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[N,V,PREP,N,ADJ,N,N,ADJ,N,NA,NA,NA,NA,NA,NA]
[V,VPRON,CONJ,N,V,VPRON,N,ADJ,ADV,N,V,PREP,N,ADJ,ADV]
[CONJ,ADJ,PREP,ADJ,N,N,V,PART,N,PART,PREP,ADJ,CONJ,ADJ,PREP]
[ADJ,N,N,V,ADJ,PREP,ADJ,N,ADV,PART,ADJ,PREP,N,N,ADV]
[CONJ,ADV,PREP,ADJ,N,V,ADJ,N,PART,N,CONJ,V,N,ADJ,N]
[PREP,N,PREP,ADJ,N,ADJ,PART,V,ADJ,N,CONJ,V,PREP,N,N]
[ADV,ADV,N,V,ADV,N,N,N,PREP,ADJ,N,NA,NA,NA,NA]
[PREP,N,CONJ,N,V,ADV,PREP,N,CONJ,N,N,ADV,ADJ,ADV,PREP]
[PREP,N,ADJ,N,PREP,N,ADJ,N,ADJ,N,N,ADJ,V,VPRON,N]
[N,PREP,N,V,N,N,PREP,N,N,ADV,PREP,N,ADJ,ADV,CONJ]
[ADJ,N,N,V,VPRON,N,N,PREP,ADJ,N,ADV,PREP,ADJ,N,V]

Polish POS text 4

[N,PART,ADJ,N,PREP,N,N,NA,NA,NA,NA,NA,NA,NA,NA]
[N,CONJ,ADJ,N,PART,ADJ,N,NA,NA,NA,NA,NA,NA,NA,NA]
[CONJ,ADJ,V,N,ADV,ADJ,CONJ,ADJ,N,V,N,ADJ,N,CONJ,ADJ]
[N,V,ADJ,PREP,N,ADJ,N,V,VPRON,PREP,N,ADJ,PREP,ADJ,N]
[ADJ,PREP,N,CONJ,N,ADV,V,VPRON,PREP,N,PREP,N,ADJ,N,ADV]
[PART,V,ADV,ADJ,N,CONJ,PREP,ADJ,ADJ,N,ADJ,PREP,N,V,ADV]
[PART,V,ADV,N,CONJ,N,CONJ,PREP,N,ADJ,ADV,PART,PREP,ADJ,N]
[N,V,ADV,PREP,ADV,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[CONJ,N,ADV,ADJ,V,ADV,ADJ,PREP,N,PREP,N,PREP,ADJ,N,N]
[N,CONJ,PART,V,VPRON,N,V,VPRON,ADJ,N,N,V,ADV,ADJ,N]
[ADV,PART,ADV,CONJ,V,ADV,N,ADJ,PREP,N,CONJ,PREP,N,ADJ,N]
[PREP,ADJ,N,PART,ADJ,N,N,PREP,N,ADJ,PREP,N,ADJ,CONJ,ADV]
[ADV,PREP,N,PREP,N,PREP,N,N,V,ADJ,ADJ,N,PREP,N,CONJ]
[PART,ADV,PREP,ADJ,N,V,ADJ,PREP,ADJ,PREP,N,N,N,V,N]
[N,ADJ,ADV,VPRON,V,V,N,ADJ,ADJ,N,PREP,ADJ,N,NA,NA]
[ADJ,N,PREP,N,N,V,N,ADJ,N,N,PREP,ADJ,CONJ,ADJ,N]
[PREP,ADJ,N,ADV,ADV,CONJ,ADJ,N,ADV,V,PREP,N,N,CONJ,PREP]
[N,V,PREP,N,ADJ,ADJ,ADV,N,N,V,VPRON,ADV,PREP,N,N]
[ADJ,N,N,ADJ,N,ADJ,PART,V,ADV,N,CONJ,N,N,ADJ,N]
[ADV,V,VPRON,CONJ,V,ADV,PREP,N,NA,NA,NA,NA,NA,NA,NA]
[PREP,N,CONJ,ADV,V,VPRON,PREP,N,N,N,V,N,N,ADV,ADJ]
[PREP,ADJ,N,ADJ,N,ADJ,N,PREP,CONJ,ADJ,V,ADJ,N,ADJ,N]

[CONJ,N,ADJ,ADJ,ADJ,N,V,ADV,ADV,ADV,PREP,ADV,ADJ,N,NA]
[ADV,CONJ,V,ADJ,N,N,N,ADJ,PREP,N,N,NA,NA,NA,NA]
[N,V,ADV,ADJ,CONJ,ADV,PREP,N,CONJ,V,N,PREP,ADJ,N,PREP]
[ADJ,N,ADJ,V,PREP,N,ADJ,N,CONJ,PART,ADV,PREP,ADJ,N,PREP]
[N,ADJ,V,CONJ,ADJ,N,N,N,N,CONJ,N,N,N,CONJ,ADJ]
[V,PREP,N,CONJ,V,V,N,ADJ,N,NA,NA,NA,NA,NA,NA]
[PREP,N,V,VPRON,CONJ,ADJ,N,PREP,N,ADJ,ADJ,PREP,ADV,PART,ADJ]
[ADV,ADV,ADV,N,N,N,V,ADJ,PREP,ADJ,ADJ,N,CONJ,V,ADV]
[V,PREP,N,N,CONJ,ADV,V,N,N,ADJ,ADV,ADV,ADJ,VPRON,PREP]
[N,PREP,N,N,V,PREP,N,ADJ,N,PREP,ADJ,VPRON,N,PREP,N]
[PART,V,N,ADV,CONJ,N,V,N,CONJ,N,N,PREP,ADJ,NA,NA]
[PREP,N,V,N,N,VPRON,ADV,ADV,CONJ,PART,N,N,ADJ,N,N]
[N,ADJ,ADJ,V,ADV,ADV,ADJ,CONJ,ADV,ADJ,ADV,ADJ,ADJ,ADJ,ADJ]
[N,ADJ,PART,V,ADV,ADV,ADJ,ADJ,ADV,ADJ,CONJ,PART,V,PREP,N]
[CONJ,N,PREP,N,V,N,PREP,N,ADJ,N,N,CONJ,CONJ,ADV,ADJ]
[ADJ,N,ADJ,ADJ,N,CONJ,N,V,N,N,CONJ,N,N,ADV,ADV]
[N,V,PREP,ADJ,N,N,PREP,ADV,ADJ,PREP,N,N,CONJ,ADV,CONJ]
[V,N,N,ADJ,ADV,ADJ,PREP,N,CONJ,ADV,PREP,ADJ,N,N,ADV]
[ADV,ADV,V,N,ADJ,PREP,ADJ,ADJ,N,N,N,ADV,PART,ADJ,PREP]
[V,ADV,N,ADJ,PREP,N,N,V,PREP,N,ADJ,PREP,ADJ,N,N]
[PREP,ADJ,N,V,ADV,N,PREP,N,PART,ADV,PREP,N,ADJ,ADJ,N]
[PART,V,N,N,ADJ,PREP,ADJ,ADJ,CONJ,ADJ,ADV,ADJ,N,N,CONJ]
[PREP,ADJ,N,ADV,N,PREP,ADV,ADJ,ADV,ADV,ADJ,N,PREP,N,N]
[V,N,PREP,N,PREP,N,ADJ,N,CONJ,V,N,ADV,ADJ,ADJ,N]
[PREP,N,ADJ,N,N,PREP,N,N,N,V,PREP,N,ADJ,N,N]
[ADV,VPRON,PREP,ADJ,CONJ,ADJ,N,N,V,VPRON,PREP,ADJ,N,PREP,
CONJ]
[ADV,N,N,PREP,ADJ,N,PREP,ADJ,ADJ,N,N,PREP,N,V,PREP]
[V,N,ADV,CONJ,ADJ,N,N,PREP,N,V,PREP,N,PREP,N,NA]

Polish POS text 5

[N,ADJ,V,VPRON,N,PREP,N,N,NA,NA,NA,NA,NA,NA,NA]
[ADV,V,N,CONJ,N,ADV,V,N,PREP,N,NA,NA,NA,NA,NA]
[V,N,ADJ,PREP,N,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[V,PREP,N,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[V,ADJ,ADJ,CONJ,ADJ,N,PREP,N,NA,NA,NA,NA,NA,NA,NA]
[V,CONJ,N,V,ADJ,N,V,N,PREP,N,CONJ,V,ADJ,N,V]
[ADV,V,PREP,N,V,N,CONJ,V,PREP,ADV,N,VPRON,N,NA,NA]
[V,N,N,V,N,V,N,CONJ,N,PREP,N,ADJ,PREP,N,V]
[V,ADV,N,N,CONJ,V,CONJ,N,V,ADJ,N,NA,NA,NA,NA]
[PART,V,N,N,ADJ,N,PREP,ADJ,N,PREP,CONJ,V,PART,N,CONJ]
[ADJ,ADJ,N,N,N,CONJ,POST,N,ADJ,N,V,N,N,N,ADJ]
[CONJ,N,ADJ,V,ADV,PREP,N,ADV,V,N,CONJ,PREP,ADJ,N,V]

[CONJ,ADV,V,VPRON,PREP,ADJ,N,V,VPRON,PREP,ADJ,N,ADJ,NA,NA]
[PREP,N,V,N,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[N,CONJ,V,N,ADJ,V,PREP,N,ADJ,ADJ,N,CONJ,ADJ,N,N]
[ADV,N,V,PART,ADV,ADJ,N,CONJ,N,ADJ,ADJ,N,NA,NA,NA]
[ADJ,CONJ,ADJ,PREP,N,ADV,PREP,ADJ,N,N,V,VPRON,PREP,N,N]
[CONJ,PREP,ADJ,N,N,V,ADV,ADV,N,NA,NA,NA,NA,NA,NA]
[N,ADJ,N,V,PREP,N,N,CONJ,N,ADJ,PREP,N,ADJ,PREP,N]
[ADV,PART,V,ADJ,ADJ,PREP,ADJ,N,ADV,ADJ,PREP,ADJ,N,ADV,ADJ]
[V,N,PREP,ADJ,N,CONJ,PART,V,N,CONJ,CONJ,PART,V,N,V]
[ADJ,ADJ,N,V,ADV,ADJ,N,ADJ,N,CONJ,N,ADJ,N,NA,NA]
[PREP,N,ADJ,N,V,N,ADV,ADV,N,NA,NA,NA,NA,NA,NA]
[V,N,CONJ,V,ADJ,N,N,N,N,N,CONJ,N,N,N,NA]
[V,ADV,ADJ,N,PREP,ADV,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[N,PART,V,ADV,N,PREP,ADJ,ADV,PREP,N,NA,NA,NA,NA,NA]
[N,ADV,PREP,ADJ,N,ADJ,N,CONJ,ADJ,N,V,NA,NA,NA,NA]
[N,V,CONJ,PREP,N,ADJ,N,ADJ,N,ADV,ADJ,PREP,N,NA,NA]
[PART,V,CONJ,PREP,N,ADJ,N,N,CONJ,V,CONJ,PREP,N,PREP,ADJ]
[ADV,V,N,N,CONJ,PREP,ADJ,CONJ,ADJ,N,V,ADV,ADJ,N,N]
[ADJ,ADJ,N,N,ADJ,N,V,VPRON,ADV,N,N,PREP,N,N,CONJ]
[ADJ,N,V,ADJ,N,CONJ,PART,V,PREP,N,PREP,N,NA,NA,NA]
[ADJ,PREP,N,N,ADJ,N,V,PREP,N,ADJ,CONJ,ADV,V,N,ADV]
[ADV,N,PREP,ADJ,N,V,PART,ADJ,N,N,CONJ,VPRON,PREP,N,V]
[ADJ,N,V,N,PREP,ADJ,N,N,NA,NA,NA,NA,NA,NA,NA]
[CONJ,V,ADV,ADJ,N,ADJ,N,PREP,ADJ,N,V,ADJ,N,CONJ,PREP]
[N,V,N,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[V,N,PREP,N,N,ADJ,CONJ,ADJ,N,ADJ,N,N,V,VPRON,ADV]
[V,N,ADV,ADJ,N,CONJ,N,CONJ,V,ADJ,N,NA,NA,NA,NA]
[ADV,N,PREP,N,ADJ,N,V,ADV,N,ADV,N,ADV,PREP,N,N]
[ADJ,N,ADV,ADJ,V,ADV,N,PREP,N,CONJ,N,V,N,N,NA]
[V,VPRON,CONJ,ADJ,ADJ,N,ADJ,N,PREP,ADJ,N,V,ADV,ADV,ADJ]
[N,V,ADV,ADJ,N,PREP,N,N,NA,NA,NA,NA,NA,NA,NA]
[N,V,CONJ,ADJ,N,ADJ,PREP,N,CONJ,PREP,N,N,ADJ,N,V]
[N,V,N,ADJ,ADJ,N,ADV,N,PREP,ADJ,N,CONJ,N,ADJ,N]
[PART,V,ADV,ADV,V,N,ADJ,N,CONJ,ADJ,N,V,ADJ,N,NA]
[V,N,CONJ,N,ADV,N,CONJ,V,PREP,N,N,N,ADV,V,PREP]
[ADV,V,ADJ,N,N,CONJ,N,PREP,N,PART,V,PREP,ADJ,N,NA]
[PREP,N,ADJ,ADV,V,PREP,N,ADJ,CONJ,ADJ,ADV,PART,ADJ,V,VPRON]
[ADJ,N,N,N,N,V,N,ADJ,ADJ,N,N,ADJ,CONJ,ADJ,NA]

Turkish

ADJ adjective
ADV adverb

ART article
CONJ conjunction
N noun
PART particle
POST postposition
V verb
NA empty space (filler to reach 15 elements)

Turkish Parts of Speech 1

[N,N,N,N,N,ADV,ADJ,ART,N,N,V,NA,NA,NA,NA]
[N,POST,N,ADJ,N,N,N,N,POST,V,CONJ,N,N,ADJ,N]
[N,ADJ,N,POST,N,ART,N,N,N,V,ADJ,N,ADJ,N,V]
[ADJ,N,N,ADJ,N,N,ADJ,N,N,ADV,ADJ,N,N,POST,N]
[ADJ,N,N,N,N,N,ADJ,ART,N,N,POST,V,NA,NA,NA]
[PART,N,N,ADJ,N,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[ADJ,N,V,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[ADV,N,N,N,ADJ,N,ADJ,N,N,V,NA,NA,NA,NA,NA]
[N,POST,N,N,N,N,N,N,N,POST,ADJ,N,ADJ,N,ADJ]
[CONJ,N,POST,ADJ,N,ADJ,N,POST,N,ADJ,N,V,NA,NA,NA]
[ADV,N,ART,N,ADJ,N,ADJ,N,ADJ,N,V,NA,NA,NA,NA]
[ADV,N,N,N,ADJ,N,ADJ,N,V,NA,NA,NA,NA,NA,NA]
[PART,N,N,N,ADJ,N,N,ADV,N,N,ADJ,N,N,ADV,ADJ]
[PART,N,N,N,V,N,N,V,ADV,V,N,N,N,N,N]
[N,POST,V,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[PART,ADV,ADV,V,N,N,V,NA,NA,NA,NA,NA,NA,NA,NA]
[ADV,V,N,N,N,N,ADJ,ADJ,CONJ,N,N,POST,ADJ,N,N]
[N,N,V,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[ADJ,N,ADJ,N,N,ADJ,N,ADJ,N,POST,N,ADJ,N,ADV,N]
[PART,N,N,N,N,N,N,CONJ,N,N,ADJ,N,N,V,NA]
[ADV,ADJ,N,N,POST,V,V,PART,PART,ADV,V,NA,NA,NA,NA]
[PART,V,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[CONJ,N,ADV,ADJ,V,ADJ,N,N,NA,NA,NA,NA,NA,NA,NA]
[CONJ,ADV,ADJ,N,N,N,N,N,V,NA,NA,NA,NA,NA,NA]
[N,N,N,N,PART,ADJ,N,V,NA,NA,NA,NA,NA,NA,NA]
[N,CONJ,ADJ,N,N,N,ADV,ADJ,V,NA,NA,NA,NA,NA,NA]
[N,CONJ,ADJ,N,N,POST,N,ADJ,N,ADJ,N,N,ADJ,ART,N]
[ADJ,N,N,N,V,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[ADJ,N,ADJ,N,ADV,ADJ,ADJ,N,N,N,V,NA,NA,NA,NA]
[CONJ,ADJ,N,ADJ,CONJ,ADJ,ADJ,N,V,CONJ,N,ADJ,N,V,NA]
[N,ADJ,N,N,N,POST,ADV,N,V,NA,NA,NA,NA,NA,NA]
[ADJ,N,N,CONJ,N,N,POST,ADJ,N,V,NA,NA,NA,NA,NA]
[N,ADJ,N,N,POST,ADV,ADV,ADJ,N,N,V,NA,NA,NA,NA]
[ADJ,N,PART,CONJ,ADJ,N,V,N,V,NA,NA,NA,NA,NA,NA]

[ADJ,ART,N,V,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[ADJ,N,ADV,V,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[ADJ,N,N,POST,ADJ,ADJ,N,POST,POST,N,V,CONJ,V,NA,NA]
[N,N,N,POST,CONJ,N,N,POST,NA,NA,NA,NA,NA,NA,NA]
[CONJ,N,POST,N,V,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[ADJ,N,N,N,POST,N,N,V,NA,NA,NA,NA,NA,NA,NA]
[CONJ,N,N,V,N,POST,ADJ,N,V,NA,NA,NA,NA,NA,NA]
[ADJ,N,POST,ADV,ADJ,V,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[CONJ,ADV,ADJ,N,N,N,N,V,CONJ,N,N,V,NA,NA,NA]
[ADJ,POST,POST,ADJ,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[N,POST,ADJ,N,N,V,N,V,ADJ,N,V,NA,NA,NA,NA]
[CONJ,ADJ,ADV,ADV,N,N,N,V,NA,NA,NA,NA,NA,NA,NA]
[N,POST,N,V,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[N,ADJ,N,POST,CONJ,N,N,V,CONJ,POST,N,POST,N,CONJ,N]
[ADJ,N,POST,N,N,N,V,NA,NA,NA,NA,NA,NA,NA,NA]
[N,N,N,ADV,ART,N,V,NA,NA,NA,NA,NA,NA,NA,NA]

Turkish Parts of Speech T 2

[N,ADJ,ADJ,N,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[ADV,N,ADJ,N,POST,N,N,N,N,ADJ,V,CONJ,ADJ,ART,N]
[N,ADV,N,POST,ADJ,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[N,ADV,N,N,V,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[N,POST,N,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[ADJ,ART,N,N,N,ADJ,N,ADV,N,N,N,N,ADJ,N,N]
[N,N,POST,N,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[ADJ,ADJ,N,ADJ,ART,N,V,NA,NA,NA,NA,NA,NA,NA,NA]
[N,ADJ,CONJ,N,N,POST,V,NA,NA,NA,NA,NA,NA,NA,NA]
[N,ADV,N,N,ADV,N,N,ADJ,N,N,V,N,N,V,V]
[CONJ,ADJ,N,N,ADV,N,N,N,ADJ,ART,N,V,POST,V,NA]
[N,N,N,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[N,ADJ,N,ADV,N,ADJ,N,ADV,N,POST,ADV,N,CONJ,N,N]
[N,N,N,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[ADJ,N,N,N,ADJ,N,N,V,POST,N,N,N,N,ADJ,N]
[ADJ,N,N,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[ADJ,N,CONJ,ADJ,ADJ,N,ADJ,N,N,V,NA,NA,NA,NA,NA]
[ADJ,CONJ,ADJ,N,POST,N,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[ADJ,ADJ,N,ADJ,POST,N,ADJ,ADJ,ADJ,V,POST,V,NA,NA,NA]
[ADJ,N,N,N,N,V,CONJ,ADJ,N,ADJ,N,POST,N,V,NA]
[ADJ,N,N,POST,N,V,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[N,N,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[N,N,CONJ,N,N,N,N,V,NA,NA,NA,NA,NA,NA,NA]
[ADJ,N,ADV,ADJ,N,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[ADJ,ADJ,ADJ,N,ADJ,POST,N,N,ADJ,N,N,N,CONJ,N,V]

[N,CONJ,ADJ,N,N,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[ADJ,N,CONJ,N,ADJ,N,N,N,ADV,ADJ,V,NA,NA,NA,NA]
[N,ADV,CONJ,ADJ,N,ADV,PART,ADJ,ART,N,PART,ADJ,N,CONJ,N]
[ADV,N,N,ADJ,N,ADV,ADV,N,V,CONJ,N,ADJ,V,NA,NA]
[N,N,V,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[ADJ,ADJ,ART,N,CONJ,N,N,ADV,N,ADJ,V,NA,NA,NA,NA]
[ADJ,N,N,N,ADJ,N,POST,N,N,N,N,ADJ,NA,NA,NA]
[ADJ,CONJ,ADJ,ART,N,POST,N,NA,NA,NA,NA,NA,NA,NA,NA]
[ADJ,N,ADJ,N,POST,ADJ,POST,V,NA,NA,NA,NA,NA,NA,NA]
[N,ADV,N,N,N,ADV,ADJ,N,N,ADV,V,NA,NA,NA,NA]
[ADV,N,CONJ,N,CONJ,POST,N,V,NA,NA,NA,NA,NA,NA,NA]
[N,ADJ,N,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[N,ADJ,N,ADJ,N,N,N,CONJ,ADJ,N,POST,N,ADJ,N,N]
[N,N,N,ADJ,N,N,N,N,ADJ,ADV,ADJ,N,CONJ,N,POST]
[ADJ,N,N,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[N,N,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[ADJ,N,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[ADJ,ADV,N,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[ADJ,ADJ,ADJ,N,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[N,N,ADJ,ADJ,N,N,N,N,POST,NA,NA,NA,NA,NA,NA]
[N,N,N,N,N,N,N,N,POST,ADJ,CONJ,ADJ,N,N,NA]
[ADJ,N,N,N,V,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[N,ADJ,ART,N,ADV,ADJ,N,CONJ,N,N,ADJ,N,N,N,V]
[ADJ,N,ADV,V,ADJ,ADJ,ART,N,N,NA,NA,NA,NA,NA,NA]
[ADJ,N,ADJ,N,ADV,N,ADV,POST,V,CONJ,ADJ,N,ADJ,ART,N]

Turkish Parts of Speech T 3

[N,N,N,N,N,N,ADJ,N,N,ADJ,ADJ,ADJ,N,N,V]
[N,N,N,N,N,N,ADV,ADJ,N,ADV,N,N,V,NA,NA]
[ART,N,N,N,N,ADJ,N,PART,N,V,NA,NA,NA,NA,NA]
[CONJ,ADJ,N,N,N,ADV,ADJ,NA,NA,NA,NA,NA,NA,NA,NA]
[N,N,N,N,N,N,ADJ,N,N,ADJ,ART,N,CONJ,ART,N]
[N,N,N,N,N,N,N,POST,ADJ,N,CONJ,N,N,ADJ,N]
[N,N,N,ADJ,N,N,N,V,POST,ADJ,ART,N,N,V,NA]
[N,N,CONJ,N,V,CONJ,N,V,NA,NA,NA,NA,NA,NA,NA]
[N,POST,ADJ,N,N,N,N,ADJ,ADJ,N,N,ADJ,ADJ,POST,ADJ]
[ADJ,N,N,N,POST,N,V,NA,NA,NA,NA,NA,NA,NA,NA]
[ADV,ADJ,N,ADJ,N,ADJ,N,N,ADJ,CONJ,ADJ,N,N,ADJ,N]
[CONJ,ADJ,POST,ADJ,N,ADV,ADJ,POST,ADJ,PART,V,NA,NA,NA,NA]
[N,PART,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[N,ADJ,ADJ,N,N,ADV,ADV,N,ART,ADJ,N,N,N,V,CONJ]
[ADJ,N,N,POST,N,PART,N,ADJ,N,ADJ,N,N,N,ADJ,N]
[N,ADJ,N,ADJ,N,POST,ADJ,N,V,N,N,CONJ,ART,PART,N]

[CONJ,N,N,ADJ,N,ADJ,N,PART,ADV,N,N,ADJ,V,NA,NA]
[N,N,N,POST,N,ADJ,ADJ,ADJ,N,N,ADJ,ART,N,V,N]
[ADJ,ADV,N,ADV,N,N,ADJ,N,POST,N,N,V,N,NA,NA]
[N,N,N,N,ADV,ADJ,N,ADJ,N,V,NA,NA,NA,NA,NA]
[N,N,ADJ,N,N,N,POST,ADJ,N,ADJ,ART,N,V,NA,NA]
[N,N,N,ADJ,N,POST,N,N,ADJ,N,ART,N,N,N,N]
[ADJ,N,N,N,N,N,N,V,NA,NA,NA,NA,NA,NA,NA]
[ADJ,N,ADJ,ADJ,N,N,ADV,N,POST,ADJ,N,ADV,N,N,POST]
[PART,ADJ,N,ADJ,ART,N,V,NA,NA,NA,NA,NA,NA,NA,NA]
[N,N,N,ADJ,POST,V,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[ADJ,N,ADJ,ART,N,N,N,ADV,N,N,N,N,N,ADJ,N]
[N,V,POST,N,N,V,N,NA,NA,NA,NA,NA,NA,NA,NA]
[N,N,POST,N,V,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[N,N,ADJ,ADJ,N,ADJ,ADJ,POST,N,ADJ,V,ADJ,ART,N,NA]
[N,ADJ,N,ADV,ADJ,ART,N,N,ADJ,N,ADJ,ADV,N,N,N]
[N,N,V,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[N,ART,ADJ,N,V,N,N,ADJ,V,ADV,N,N,ADJ,ART,N]
[N,ADV,N,N,ADJ,N,N,N,ADJ,N,N,ADJ,N,N,ADJ]
[ADJ,N,N,N,N,N,N,N,N,POST,ADJ,N,ADJ,N,ADJ]
[CONJ,N,N,N,CONJ,ADJ,N,N,CONJ,N,N,ADJ,POST,ADV,N]
[ADV,CONJ,ADJ,N,N,ADV,ADV,ADJ,N,N,ADV,ADJ,V,NA,NA]
[CONJ,PART,N,N,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[ART,N,N,PART,ADJ,N,ADJ,N,N,N,NA,NA,NA,NA,NA]
[N,N,N,ADJ,N,N,N,N,POST,ART,N,N,POST,V,CONJ]
[N,ADV,N,PART,N,N,N,PART,N,N,V,PART,POST,N,ART]
[N,ADJ,N,N,ADV,N,ADJ,N,CONJ,N,N,ADJ,N,POST,ADV]
[N,N,ADV,N,N,ADJ,N,N,N,N,V,NA,NA,NA,NA]
[N,N,ADJ,ADJ,N,N,ADJ,N,N,ADJ,N,V,N,N,NA]
[PART,ADJ,N,ART,N,ADJ,N,POST,N,POST,N,PART,CONJ,N,N]
[ADV,ADJ,N,N,N,N,N,N,V,ADV,ADJ,N,V,ADJ,N]
[ADJ,N,N,ADJ,N,N,N,ADV,ADV,ADJ,N,N,N,V,NA]
[N,ADJ,ADJ,N,N,ADJ,N,PART,N,N,N,N,N,ADV,V]
[N,ADJ,N,POST,N,PART,N,V,CONJ,N,POST,N,N,N,N]
[N,N,ADV,ADJ,N,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]

Turkish Parts of Speech T 4

[N,ADV,N,N,ADJ,N,CONJ,N,N,NA,NA,NA,NA,NA,NA]
[N,POST,N,CONJ,N,N,POST,N,ADJ,ADJ,N,ADV,N,POST,ADJ]
[ADJ,ADJ,N,N,POST,ADJ,N,ADJ,ADJ,N,POST,ART,N,N,POST]
[N,ADV,ART,N,N,N,ADJ,ADJ,N,CONJ,N,N,N,CONJ,N]
[N,N,ADJ,N,N,N,N,N,N,N,N,POST,ADJ,N,ADJ]
[N,ADJ,N,ART,N,ADJ,N,ADV,N,NA,NA,NA,NA,NA,NA]
[N,N,ADV,ADV,N,N,POST,N,ADJ,N,V,NA,NA,NA,NA]

[N,N,ADJ,N,CONJ,N,N,N,N,N,N,N,N,CONJ,ADJ]
[ADJ,N,ADJ,N,N,POST,ADJ,ADJ,N,N,N,N,V,NA,NA]
[N,N,N,N,N,N,N,N,N,N,N,N,CONJ,ADJ,POST,N]
[N,N,N,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[N,CONJ,N,PART,N,N,N,POST,V,NA,NA,NA,NA,NA,NA]
[ADJ,N,N,N,PART,N,N,NA,NA,NA,NA,NA,NA,NA,NA]
[N,N,ADV,ADJ,N,N,CONJ,N,POST,ADJ,N,N,N,ADJ,N]
[ADJ,N,N,ADJ,N,N,N,ADJ,ART,N,N,N,N,N,CONJ]
[N,POST,N,N,POST,N,ART,N,V,N,POST,N,N,POST,N]
[N,N,ADJ,N,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[N,ADJ,N,N,ADJ,N,ADJ,V,PART,ADJ,N,N,ADJ,N,ADV]
[N,ADJ,N,ADJ,ART,N,PART,PART,ART,N,ADV,N,ADV,POST,N]
[ADV,ART,N,CONJ,PART,N,ADJ,ADV,N,N,NA,NA,NA,NA,NA]
[ADV,ADJ,ART,N,N,N,N,ADJ,N,N,N,V,NA,NA,NA]
[ADV,POST,N,N,ADJ,CONJ,PART,ADJ,N,N,POST,ADJ,N,ADJ,V]
[ADJ,N,N,ADJ,ADJ,N,N,ADJ,N,POST,N,NA,NA,NA,NA]
[N,N,POST,N,ADJ,PART,ADJ,N,N,PART,N,N,N,ADJ,N]
[CONJ,N,POST,N,N,N,POST,N,ADJ,N,ADJ,N,N,N,CONJ]
[N,CONJ,ADJ,ADJ,N,POST,ADJ,N,N,PART,V,NA,NA,NA,NA]
[N,N,N,V,CONJ,N,N,N,N,ADJ,ADJ,N,ADJ,CONJ,N]
[N,N,POST,N,ADJ,N,N,ADJ,PART,ADJ,N,ADJ,ADJ,N,CONJ]
[ADJ,N,ART,N,ADJ,N,N,ADV,N,CONJ,ADJ,PART,N,ART,N]
[ADJ,N,ART,N,N,ADJ,ADJ,N,N,N,N,N,NA,NA,NA]
[N,N,N,ADJ,ADJ,N,ADJ,ADJ,N,N,ADJ,ART,N,POST,N]
[N,N,ADJ,N,N,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[N,ADJ,N,ART,N,ADJ,N,N,N,N,V,NA,NA,NA,NA]
[N,ADJ,ART,N,V,CONJ,N,POST,ADJ,ADJ,N,ADJ,N,V,NA]
[N,N,CONJ,N,N,N,ADJ,N,N,ADJ,ART,N,V,NA,NA]
[N,CONJ,N,POST,ADJ,ART,N,V,NA,NA,NA,NA,NA,NA,NA]
[ADJ,N,ADV,N,V,N,POST,N,N,N,N,N,N,ADJ,ART]
[ADJ,N,ADJ,N,POST,N,N,N,V,NA,NA,NA,NA,NA,NA]
[ADJ,ADJ,ART,N,PART,N,ADJ,N,N,ADJ,ADJ,N,V,NA,NA]
[ADV,ADJ,ADJ,N,ADJ,N,N,N,N,CONJ,ADJ,N,V,NA,NA]
[N,ADV,N,N,POST,PART,ADV,V,NA,NA,NA,NA,NA,NA,NA]
[N,N,V,N,N,N,N,N,POST,ADJ,N,ADV,ADJ,N,N]
[N,N,ADV,ADJ,ADV,PART,N,ADJ,N,ADJ,N,V,NA,NA,NA]
[N,N,ADJ,N,N,POST,ADJ,N,N,N,CONJ,ADJ,N,N,ADV]
[CONJ,ADJ,N,ADJ,N,CONJ,N,N,N,N,V,N,N,ADJ,ADJ]
[N,N,N,ART,ADJ,ADJ,N,PART,PART,ADJ,N,N,N,ADJ,N]
[ADV,PART,ADJ,N,N,V,PART,N,N,PART,N,N,V,NA,NA]
[N,N,N,N,N,N,N,ADJ,N,N,N,V,NA,NA,NA]
[N,PART,N,N,N,POST,N,ADJ,ART,N,N,V,NA,NA,NA]
[N,PART,N,N,N,ART,N,N,N,NA,NA,NA,NA,NA,NA]

Turkish Parts of Speech T 5

[N,N,ADJ,N,N,N,POST,ADJ,N,PART,ADJ,ADJ,N,POST,V]
[ADJ,ART,N,N,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[N,PART,V,N,N,ADJ,N,ADJ,ADJ,N,N,V,N,N,N]
[ADJ,N,N,N,N,ADJ,CONJ,ADJ,N,N,N,POST,N,N,N]
[ADJ,N,N,ADJ,CONJ,ADJ,N,N,V,CONJ,N,ADJ,N,POST,N]
[N,ADJ,N,N,N,N,N,N,POST,V,NA,NA,NA,NA,NA]
[N,N,ADJ,N,N,N,ADJ,N,N,N,N,CONJ,N,ADJ,N]
[N,N,N,N,ADJ,N,N,N,ADJ,N,N,N,ADJ,N,V]
[N,ADJ,N,N,N,N,N,N,ADJ,ADV,N,N,N,V,CONJ]
[N,N,N,ADJ,N,N,N,POST,N,N,N,V,NA,NA,NA]
[N,ADV,N,POST,N,V,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[ADJ,N,ADV,N,N,N,N,POST,N,NA,NA,NA,NA,NA,NA]
[ADJ,CONJ,N,N,ADJ,N,CONJ,N,N,N,N,V,NA,NA,NA]
[ART,ADJ,N,N,N,ADV,ADJ,NA,NA,NA,NA,NA,NA,NA,NA]
[ADJ,N,N,N,V,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[ADJ,N,N,CONJ,ADJ,N,POST,N,N,V,NA,NA,NA,NA,NA]
[CONJ,N,N,N,ADJ,N,ADJ,N,ART,N,N,N,CONJ,ADJ,N]
[N,N,N,CONJ,N,N,N,NA,NA,NA,NA,NA,NA,NA,NA]
[ADJ,ART,N,N,N,N,N,N,N,ADJ,ADJ,N,CONJ,N,N]
[ADJ,N,N,ADV,ADJ,N,ADJ,N,ADJ,N,POST,N,CONJ,N,ADJ]
[N,N,ADJ,N,N,POST,ADJ,ART,N,V,NA,NA,NA,NA,NA]
[CONJ,N,ADJ,N,N,ADJ,N,CONJ,N,ADJ,N,ADJ,ART,N,V]
[N,ADV,ADJ,ADJ,N,CONJ,N,N,ADJ,N,V,NA,NA,NA,NA]
[ADJ,N,CONJ,ADJ,N,CONJ,PART,N,N,N,ADJ,N,V,NA,NA]
[ADV,N,N,N,N,ADJ,N,ADJ,N,ADJ,ART,N,N,N,ADJ]
[N,CONJ,N,ADJ,N,N,N,N,ADJ,N,ADJ,ART,N,N,N]
[N,ADJ,ART,N,N,N,POST,ART,N,CONJ,N,POST,N,N,N]
[ADV,ADJ,POST,ADJ,ADJ,N,ADV,V,CONJ,N,N,ADJ,N,N,N]
[N,N,CONJ,N,N,N,N,CONJ,ADJ,N,ADJ,ADV,N,N,ADJ]
[N,N,N,N,POST,ADV,ADV,ADJ,POST,ADJ,N,N,V,NA,NA]
[ADV,N,N,V,CONJ,N,N,ART,N,N,N,ADJ,ADJ,ADJ,ART]
[N,N,POST,ADJ,N,N,N,CONJ,N,N,N,N,ADJ,ADJ,N]
[ADJ,N,N,ADJ,N,ADJ,N,ADJ,N,NA,NA,NA,NA,NA,NA]
[N,ADJ,N,ADJ,N,N,N,ADJ,N,ADJ,N,V,NA,NA,NA]
[N,POST,N,ADJ,N,N,CONJ,ADJ,ADJ,N,ADJ,N,V,NA,NA]
[ADJ,N,ADJ,N,N,ADJ,N,N,N,NA,NA,NA,NA,NA,NA]
[N,N,N,ADJ,N,V,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[N,N,ADJ,PART,ADJ,N,ADV,ADJ,N,N,ADJ,ADJ,N,ADJ,N]
[N,N,V,N,N,V,CONJ,ADJ,N,ART,ADJ,N,ADV,N,ADV]
[N,CONJ,N,N,N,N,N,ADJ,N,ADV,ADJ,V,NA,NA,NA]
[N,POST,N,V,PART,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[ADV,V,CONJ,N,ADJ,V,ADJ,N,CONJ,N,POST,NA,NA,NA,NA]

[ADJ,ADJ,N,POST,N,N,N,N,ADJ,ART,N,N,ADJ,V,NA]
[N,ADJ,N,POST,ART,N,V,CONJ,N,N,V,NA,NA,NA,NA]
[PART,CONJ,N,ART,N,V,CONJ,N,ADJ,N,ADV,N,N,V,NA]
[N,ADJ,N,ADJ,ART,N,N,N,CONJ,ADJ,ART,N,N,ADJ,N]
[ADJ,N,ADV,ADJ,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[PART,ADJ,N,ADV,V,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA]
[N,CONJ,N,ADJ,CONJ,N,N,N,N,POST,V,NA,NA,NA,NA]
[ADJ,N,V,CONJ,N,N,N,V,NA,NA,NA,NA,NA,NA,NA]

Turkish Word length T 1

[2,4,3,3,2,4,2,2,1,2,4,2]
[3,2,2,3,3,4,2,2,2,3,4,2,2,3,2,2,2,4,2,2,1,2,3]
[3,3,3,4,2,1,1,2,2,3,2,1,1,1,1,2,3,2,1,2,1,3,2]
[1,1,3,3,4,3,2,3,4,4,2,2,2,6,2,1,2]
[2,2,2,4,2,3,3,1,3,4,1,4]
[2,3,3,4,1,3]
[2,2,2]
[1,3,4,3,1,2,3,2,5,4]
[2,2,2,2,4,3,5,1,3,3,2,4,2,2,2,2,4,2,2,2,2,4,3]
[2,2,3,2,2,2,4,2,2,1,2,4,3]
[3,1,1,2,2,2,2,1,2,2,2,3,2,1,2,2,2,1,1,2,2,2,3,4]
[3,2,3,2,1,1,1,1,2,3,2,1,2,1,1,3,2]
[2,2,3,2,1,1,3,3,5,3,2,3,3,4,2,2,2,6,2,1,2]
[2,2,2,4,6,2,3,3,2,3,2,3,3,2,3,2,4]
[4,2,2]
[2,2,2,2,3,2,4]
[2,5,3,1,4,2,2,3,1,2,3,3,1,4,3,1,2,1,3,1,5,4,4,3,4]
[2,4,3]
[1,3,3,3,3,2,2,2,4,2,2,2,1,4,4,3,2,1,2,4,2]
[2,5,2,3,4,4,3,1,3,6,1,3,2,4]
[2,4,1,3,3,3,5,1,2,2,4]
[2,7]
[2,6,1,3,3,1,2,4]
[2,1,2,2,4,6,1,1,3,2]
[2,1,3,4,1,2,4,2]
[2,1,2,3,4,4,1,2,3]
[2,1,2,3,3,4,5,1,3,2,2,3,3,1,6,6,4,4,4,1,3,1,2,4,1,1,3,2,4,4,3,2,5,4]
[1,2,1,2,2,3,2,3,5,4,2,2]
[1,2,1,2,2,3,3,2,3,2,2,3,3,3,3]
[2,4,3,5,1,4,4,3,4,1,3,2,3,3]
[5,5,4,3,4,2,3,2,2]
[1,3,2,1,0,4,1,4,2,2]
[3,1,2,5,3,4,1,2,2,3,2]

[1,3,2,1,2,4,4,3,4]
[1,1,3,5]
[1,2,3,4]
[3,2,2,1,1,2,1,2,2,0,2,1,3,5,2,3]
[3,2,4,2,2,3,3,2]
[2,2,1,4,4]
[2,2,3,3,1,2,2,3]
[2,1,3,2,4,2,4,2,4]
[1,2,1,1,1,4]
[2,3,1,1,2,2,3,2,4,1,2,3,3]
[4,2,1,2]
[1,2,2,3,3,4,5,4,2,3,6]
[2,3,3,4,4,3,3,5]
[2,2,3,4]
[4,2,4,3,1,3,1,4,1,1,4,3,1,1,2,2,2,2,4,4]
[1,2,1,2,2,3,3]
[5,2,3,2,1,1,3]

Turkish Word length T 2

[1,1,2,3]
[2,4,2,4,3,4,2,4,3,2,4,2,2,1,2,2,1,1,1,1,3]
[1,2,1,2,3]
[3,2,3,3,3]
[5,3,7]
[2,1,3,3,2,4,4,5,5,3,2,3,2,5,3,3]
[4,2,3,6]
[1,1,2,2,1,2,5]
[3,5,2,4,2,3,7]
[3,3,2,2,3,2,2,2,3,4,5,4,2,3,7,1,1,4,2,2]
[2,1,2,2,3,3,3,3,3,1,3,3,1,4]
[2,4,6]
[2,5,2,7,2,5,3,7,6,3,8,4,2,6,7,2,2,2,2,4,5,4,3,3]
[4,6,6]
[2,7,3,4,3,3,4,3,1,3,4,2,5,3,2,3,2]
[2,5,6]
[6,2,2,6,2,3,3,5,6,5]
[3,2,3,4,3,5]
[4,2,5,5,1,3,3,2,3,3,3,3]
[1,3,5,2,4,2,1,1,2,2,5,2,4,3]
[2,5,5,1,2,3]
[4,8]
[4,5,1,3,3,2,4,5]
[1,4,4,6,3]

[2,4,2,2,3,2,4,6,2,6,4,3,1,4,2]
[1,2,2,2,6]
[3,4,2,2,2,4,3,5,2,2,2]
[1,4,2,2,2,2,1,2,1,2,2,2,3,2,5,2,2,5]
[3,4,3,3,4,1,2,1,2,1,4,1,2]
[2,4,3]
[2,3,1,2,2,2,2,4,5,2,3]
[1,3,3,3,3,4,2,2,1,2,5,4,4]
[3,2,3,1,2,3,5]
[1,2,3,6,2,3,1,4]
[4,4,3,2,5,4,5,3,3,1,3]
[2,2,1,4,1,1,4,2]
[5,2,4]
[5,4,3,5,3,2,5,1,1,2,4,5,4,2,6,4]
[5,2,3,3,4,6,6,5,3,3,4,4,1,2,3,5,6]
[4,5,6]
[7,7]
[3,6]
[2,3,6]
[2,1,3,5]
[4,3,3,3,4,3,2,4,2]
[3,3,3,2,4,2,2,3,2,3,1,2,4,6]
[2,2,3,1,2]
[3,2,1,2,2,2,7,2,3,6,3,4,4,1,4]
[2,5,4,7,1,2,1,3,3]
[2,2,2,5,4,3,3,1,7,2,1,4,2,1,3,2,2]

Turkish Word length T 3

[4,3,2,2,1,2,4,2,3,4,4,1,2,5,1,3,2,3]
[4,6,4,2,2,5,3,5,3,3,2,4,3]
[1,4,4,3,2,3,3,2,1,3]
[2,1,3,2,6,1,2]
[4,1,2,4,2,3,1,2,3,3,1,3,1,1,2,2,4,2,2,6,4,5]
[5,3,4,2,3,3,1,2,2,5,1,3,3,2,3,6,1,1,2,3,1,2]
[2,4,6,2,3,4,1,2,2,2,1,4,3,3]
[3,2,1,3,1,2,3,1]
[3,3,2,4,2,5,3,2,1,3,4,3,1,1,2,4,1]
[1,4,6,2,3,2,3]
[3,2,2,5,5,1,2,4,3,1,1,3,4,3,4,3,2,3,6,4,1,1,4,1,2,3]
[2,4,2,3,3,2,4,2,3,1,5]
[2,1]
[4,2,2,1,4,4,2,2,2,1,2,2,3,3,3,1,3,2,2,3,3,4]
[1,3,5,4,3,1,1,3,1,3,2,3,2,3,3,1,2,3,7,1,5,3]

[3,6,4,2,4,2,3,3,5,4,4,1,1,1,2,3,3,4,2,3,4]
[2,3,2,1,2,4,1,1,3,5,2,2,2]
[4,6,4,3,6,4,1,2,3,1,2,6,3,1,1,3,4]
[2,5,3,2,4,4,7,4,3,3,2,3,2]
[3,3,4,4,3,4,4,5,6,4]
[1,3,2,3,3,5,2,3,2,1,4,2,1,1,1]
[6,3,5,3,4,1,2,1,2,3,3,2,2,2,3,1,2,1,3,3,2,4,3,4,1,3,3,4,3]
[1,4,2,1,3,3,2,4,4]
[4,3,4,2,3,4,3,3,2,3,2,2,2,2,2,3,6,2,4,1,5,3,3,1,1]
[2,5,1,3,1,2,3]
[1,4,1,4,4,2,4]
[3,4,4,1,2,4,3,3,4,3,3,3,4,5,3,4,2]
[4,3,2,4,2,3,2]
[4,3,3,4,4]
[3,2,3,2,4,2,4,2,3,2,3,5,1,2]
[4,3,3,4,2,1,2,2,2,3,2,2,3,3,4,3,4,4]
[4,2,3]
[4,1,3,6,5,4,2,2,3,3,3,4,3,1,3,2,1,3,2,1,3,1,1,3,2,1]
[1,2,3,4,7,3,2,5,5,1,3,4,3,2,3,3,2,1,2,1,4,4]
[4,3,1,3,2,4,5,3,4,2,2,3,2,4,4,3,5]
[2,2,2,2,3,2,3,4,2,4,4,4,2,4,2,5]
[2,1,1,4,2,3,5,3,2,2,3,2,2,1,5]
[1,1,2,4]
[1,3,5,1,2,2,3,2,2,3]
[3,4,3,3,3,1,2,1,2,3,3,3,5,1,1,4,4,4,1,3,2,2,3,1,2,4,1,3]
[1,2,1,1,2,5,5,1,1,2,4,1,1,4,3,2,2,4,1,4,4,4]
[2,4,3,3,2,2,4,3,1,2,4,3,3,2,3,3]
[4,3,2,4,4,7,4,3,3,2,3]
[1,2,1,2,2,3,3,3,3,5,1,5,1,2,3,4,2,3,6]
[3,1,2,1,2,2,4,1,3,4,1,2,4,2,3,2,1,1,2,4,4,1,1,3]
[3,4,2,1,2,4,2,3,3,3,6,4,1,2,2,4,2,3,3,2,4,3,6,4,1,1,3,5,4,2,4,1,3,3,2]
[1,3,2,3,1,1,4,3,2,2,2,3,3,4,5]
[2,3,2,2,2,4,1,2,1,2,4,4,3,4,2,3,3]
[2,1,2,4,2,4,4,2,1,3,1,1,4,1,4,4,1,3,4,5,1,4,6,1,2]
[5,2,4,1,2,4]

Turkish Word length T 4

[4,1,4,3,3,2,1,4,1]
[2,3,3,1,3,2,3,3,3,1,1,4,3,2,4,3,4,4]
[6,1,2,4,4,3,3,6,3,4,4,1,3,5,2,1,5]
[6,4,1,3,2,4,4,2,3,1,4,2,3,4,1,5,3,4,1,4,4,1,2,4,4,3,2,4,5,2,2]
[2,4,2,3,5,2,4,2,2,3,3,2,3,5,5,3,5,1,4]
[2,3,6,1,2,5,5,3,2]

[3,2,3,3,4,3,4,3,3,4,5]
[4,4,2,5,2,3,3,5,3,3,5,3,4,1,3,3,4,4,2,3,4,1,2,4,4,4,4,2,3,3,2]
[1,3,2,3,5,2,2,2,5,3,2,4,5]
[3,4,2,1,2,2,2,2,4,3,4,5,1,3,1,3,3,4,2,5,1,1,2,2,2,4,2,3,3,1,1,2,1,2,4,4,5]
[3,5,2,1]
[4,1,4,2,1,1,2,1,1,1,2,4,2,6]
[1,6,4,3,1,4,4]
[1,8,1,4,2,4,1,3,3,2,5,2,4,3,4,1,3,2,2,2,5,3]
[5,2,4,5,5,5,3,1,2,2,1,4,1,1,2,6,1,3,3,5]
[5,2,2,2,2,3,1,1,3,3,2,2,2,2,2,5,1,1,2,2,2,2,2,2,2,4,5,3,4]
[5,4,1,3]
[2,3,4,4,1,3,1,3,2,1,3,5,1,3,2,4,2,5,3,3,3,2]
[5,2,5,3,1,3,1,2,1,2,2,4,1,3,2,3,1,2,3]
[3,1,3,1,1,3,3,3,2,2]
[3,2,1,3,1,4,2,3,5,2,5,5]
[2,3,4,4,2,1,1,1,3,2,3,1,1,2,3,4,5]
[1,3,4,3,1,3,4,2,3,3,6]
[5,4,4,3,3,1,3,4,3,1,3,5,5,2,3]
[2,5,2,2,4,1,5,3,2,4,2,3,3,4,1,1,3,3,2,1,1,2,3,3,3,2,1,4,2,2,4,4,4,5]
[5,1,2,5,4,3,3,3,4,1,2]
[5,3,2,4,2,5,3,1,3,3,3,4,2,1,3,2,1,3,2,1,4,3,1,3,2,1,7,2,3,1,2,6,5]
[3,5,3,2,7,5,4,3,1,3,1,5,2,3,1,1,2,5,1,3,3,4,4]
[1,2,1,2,2,2,4,3,3,1,1,1,3,1,4,2,2,4,4]
[1,4,1,3,4,2,1,1,1,5,3,3]
[3,2,4,2,1,3,3,1,2,4,1,1,2,2,4,3]
[5,3,3,2,1,4]
[3,2,3,1,3,2,2,3,3,3,2]
[4,3,1,2,2,1,4,2,2,3,3,1,3,3]
[1,1,1,2,4,5,2,4,4,1,1,3,4]
[1,1,1,4,2,1,1,2]
[1,2,2,4,3,5,2,2,4,2,5,4,5,3,1,2,3,4,2,1,2,3,2,2,3,3,2,2,3,6]
[1,2,4,1,4,2,3,4,3]
[2,2,1,2,1,1,2,3,3,1,2,3,5]
[3,2,2,3,2,3,3,5,5,1,2,3,6]
[4,4,5,4,4,1,3,5]
[3,2,3,1,2,1,2,2,2,3,1,4,2,3,5,3,1,2,3,3,6]
[3,3,3,3,4,2,3,1,1,2,2,3]
[2,3,2,3,1,2,1,2,2,4,4,3,3,4,1,2,5,5,3,5,4]
[2,2,3,2,3,1,2,3,2,2,2,2,4,5,3,3,3]
[3,2,5,1,3,2,4,1,1,2,3,3,2,5,3,4]
[3,1,2,3,5,2,1,2,4,1,4,3,3]
[2,4,1,3,2,4,3,3,4,3,2,3]
[1,1,3,3,4,5,4,2,1,2,2,4]
[4,1,3,2,5,1,3,1,4]

Turkish Word length T 5

[3,3,3,2,3,2,4,2,1,1,1,3,1,2,3,4,3,3]
[3,1,4,1]
[1,1,2,3,4,1,2,1,3,3,2,3,2,4,4,1,4,5,2,3,2,3,5,4,1,3,1,3,10,2,4]
[4,2,2,2,4,4,3,1,3,3,4,4,2,2,5,4,3]
[1,1,3,3,1,4,3,5,3,1,2,1,3,2,2,3,2,2,4,4,2,2,3,3,3,2,2,5,3]
[4,2,1,2,1,1,2,4,6,5,3,3,5,3,4]
[3,4,3,2,5,2,2,3,3,2,1,4,3,6,1,2,3,4,4,4,4]
[2,4,3,6,3,3,2,5,3,3,2,5,3,3,2,3,4,7,4]
[1,1,5,3,2,4,3,2,1,4,3,2,4,4,3,3,3,1,1,2,6,3,4]
[2,3,4,2,4,2,5,2,4,2,7,3,2]
[4,3,4,2,6,5]
[2,2,4,3,4,2,4,2,4]
[3,1,3,3,4,2,1,6,2,2,7,4,4]
[1,3,3,4,4,2,2,4]
[3,3,2,1,5]
[1,3,4,2,3,2,4,1,1,1]
[2,2,2,4,3,2,4,4,4,2,1,1,3,1,5,3,3,1,1,5,3,4,3]
[5,2,6,2,4,7,2]
[2,1,3,6,2,4,1,5,1,3,3,4,2,2,4,1,3,3,5,4,4,4]
[3,3,3,4,3,4,5,2,3,3,4,1,1,2,3,5,2,1,3,4,3,3,4,3,5,3,2]
[6,4,3,3,4,2,2,1,3,4]
[2,1,4,3,4,3,2,1,5,4,5,2,1,5,3]
[4,2,2,3,4,1,6,2,6,5,8]
[1,3,1,3,2,1,1,6,6,2,2,3,4]
[3,2,4,1,5,3,2,3,3,2,1,3,3,5,7,4,3,3,3,3,3,2,2,1,2,5]
[1,1,3,5,4,4,3,3,3,2,3,1,2,1,2,4,1,2,5,3,2,4]
[2,2,1,5,3,3,2,1,4,1,4,3,2,2,2,2,3,5]
[4,1,2,3,2,3,3,6,1,2,5,2,4,7,4,5,2,1,4]
[3,5,1,3,3,4,4,1,3,3,3,3,3,6,4,1,4]
[4,4,5,3,3,3,4,7,2,3,4,6,3]
[2,4,4,4,1,4,2,1,3,6,6,3,3,3,1,1,3]
[5,3,3,4,3,4,4,1,3,4,2,4,1,2,4,5,2,1,3,1,5,2,4,3]
[1,2,3,2,4,4,4,4,2]
[3,2,4,2,2,3,5,1,3,4,3,5]
[5,2,4,3,4,6,2,3,2,4,3,3,4]
[1,2,1,4,7,3,2,4,1]
[4,2,2,2,3,1,1,2]
[2,4,4,1,2,2,2,1,3,3,2,3,3,1,5,2,6]
[3,3,1,3,4,1,1,3,4,1,1,1,2,3,4,4,2,4]
[1,1,2,6,3,2,6,6,4,1,4,5]
[2,3,6,4,2]

[4,5,1,4,4,4,2,4,1,7,2]
[4,2,1,3,4,2,2,6,2,1,4,3,2,3]
[4,4,3,2,1,2,3,2,3,6,4]
[2,1,3,1,3,2,1,5,1,1,1,1,2,3,4,4]
[1,1,2,2,1,3,3,4,1,2,1,3,3,1,1,2,3]
[1,2,1,3]
[2,4,2,3,2]
[3,1,3,4,1,1,2,6,3,2,3]
[3,6,3,1,4,1,6,3]

Persian texts

Text 1

a) POS sentence-wise:

[ADV,P,N,N,P,N,N,N,A,V,ADV,P,N,DET,N,CON,N,DET,N,P,N,N,P,PRO,N,V]
[ADV,P,N,N,N,CON,ADV,N,A,AUX,P,N,N,V,CON,PRO,P,N,N,A,V,V,CON,P,N,N,N,A,P,N,CON,N,V]
[ADV,CON,N,N,P,N,A,CON,A,N,P,N,A,CON,A,P,N,N,N,P,N,V]
[P,DET,N,N,ADV,P,N,N,V,CON,A,N,CON,N,ADV,N,P,N,CON,N,N,A,V,CON,P,PRO,V]
[ADV,AUX,V,P,DET,N,N,A,A,V,CON,N,CON,N,VCON,N]
[N,N,N,RA,P,PRO,N,V,CON,P,N,N,N,A,N,RA,V]
[ADV,P,DET,N,N,ADV,P,N,N,V,N,AUX,CON,N,P,N,A,N,A,V,CON,DET,N,P,N,N,A,A,V,CON,N,P,N,N,N,P,N,ADV,A,V]
[ADV,DET,N,RA,AUX,N,A,CON,A,N,N,V,CON,N,N,N,RA,P,PRO,N,V,A,V]
[P,N,PRO,ADV,ADV,N,N,A,V,N,CON,N,N,P,N,V,CON,ADV,N,A,A,A,P,N,N,RA,V]
[AUX,V,N,PRO,P,N,P,N,N,N,A,CON,A,N,V]
[CON,A,N,A,N,P,N,N,N,CON,N,N,V]
[P,A,N,N,CON,ADV,N,N,N,N,N,A,CON,N,P,N,A,N,P,N,A,V]
[P,A,N,A,A,N,N,A,RA,P,PRO,V,CON,P,N,P,N,PRO,N,A,P,PRO,V]
[A,V,CON,N,N,CON,N,P,N,P,N,V,CON,P,N,V,CON,N,A,N,CON,N,A,RA,V]
[N,N,N,CON,P,N,N,P,PRO,V,P,N,A,V]
[PRO,N,CON,P,N,N,N,CON,P,N,N,N,A,A,CON,N,V,CON,P,N,N,N,V]
[P,N,PRO,P,N,N,N,A,N,A,P,PRO,N,CON,N,V,CON,DET,N,P,N,V,CON,P,N,N,A,A,V]
[ADV,N,N,N,A,N,A,N,V,CON,ADV,P,N,P,N,A,CON,PN,P,PRO,V]
[PRO,P,N,A,V,N,P,N,V,CON,N,N,N,A,V]
[P,N,N,N,P,N,N,A,V,CON,AUX,N,A,P,N,CON,N,N,CON,ADV,N,A,V,CON,CON,PRO,P

,N,CON,N,CON,N,A,CON,A,N,V]
[ADV,N,ADV,N,CON,N,A,AUX,P,N,CON,N,N,N,V]
[ADV,N,P,N,N,A,CON,N,N,A,P,N,N,P,N,CON,N,N,N,V,CON,N,N,RA,P,N,N,N,V,CON,N,N,N,RA,P,N,PRO,V]
[ADV,N,N,N,P,N,A,N,A,A,V,ADV,N,P,N,P,N,N,RA,ADV,V]
[ADV,DET,N,P,N,V,CON,P,N,N,N,CON,N,PRO,N,A,V,CON,PRO,RA,CON,AUX,N,N,V,A,P,N,CON,N,V]
[PRO,V,ADV,N,N,N,P,N,N,CON,N,A,V,CON,A,N,N,N,V,CON,AUX,N,A,N,P,N,N,V]
[N,V,N,P,N,N,CON,ADV,A,N,V,ADV,P,DET,N,CON,N,N,PRO,A,P,N,A,V]
[N,A,N,N,P,N,A,V,CON,N,A,N,P,N,CON,N,P,N,N,N,PRO,ADV,N,N,V]
[P,N,N,CON,N,N,A,N,V,CON,N,N,P,N,A,N,A,CON,N,A,P,N,P,N,PRO,N,A,V,CON,N,ADV,P,N,P,DET,A,N,P,N,N,N,CON,N,V]
[ADV,N,PRO,N,A,V,CON,N,P,N,N,CON,N,A,P,N,CON,N,V]
[N,N,A,N,P,N,N,CON,N,P,N,N,V]

b) Word length sentence-wise:

[2,1,3,2,3,2,3,3,3,2,2,3,2,1,2,1,2,1,3,1,1,3,1,1,3,3]
[4,1,3,3,3,1,5,5,3,2,1,3,3,8,1,2,1,2,2,2,1,5,3,1,1,3,2,3,1,5,1,3,4]
[3,5,3,2,1,3,3,1,4,2,1,4,4,1,4,4,4,2,5,4,3,5]
[1,1,3,3,3,1,1,2,1,2,3,3,1,4,4,3,3,3,1,2,2,4,6,1,1,1,6]
[5,3,1,1,1,3,3,3,2,5,1,3,1,4,5,1,1]
[3,2,5,1,1,2,3,5,1,1,3,2,3,2,3,1,5]
[3,1,1,3,2,5,3,5,2,2,2,4,3,1,2,7,5,4,6,1,1,2,1,1,2,3,4,1,1,3,1,2,4,3,1,5,1,3,1]
[3,1,3,1,3,3,4,1,2,3,3,5,1,2,2,2,1,1,2,3,2,1,4]
[1,3,1,1,3,3,2,8,3,3,1,3,3,1,3,1,1,3,4,2,3,3,1,1,3,1,5]
[2,1,2,2,1,2,1,2,2,2,4,1,4,3,1]
[4,1,2,3,3,1,3,2,4,1,3,3,1]
[1,2,4,4,1,3,3,1,3,3,2,3,1,2,1,2,4,2,5,4,3,1]
[1,1,3,2,2,2,2,3,1,1,1,3,1,4,2,1,4,1,4,3,1,1,5]
[2,1,1,3,2,1,3,3,3,4,8,1,1,1,3,3,1,3,3,3,1,3,3,1,7]
[6,3,3,1,1,2,5,1,2,3,3,1,2,1,1]
[2,6,1,1,2,4,6,2,1,2,4,4,4,5,1,2,1,1,1,3,3,5,3]
[1,3,2,1,4,3,3,3,3,2,3,2,2,1,3,3,2,1,3,1,6,4,1,3,4,2,3,3,2]
[4,3,2,3,3,3,3,2,6,3,2,4,4,1,1,3,1,4,3,1,1,5]
[2,1,3,2,2,2,1,4,1,1,4,3,3,4,1]

1,3,5,3,1,3,3,3,3,1,4,1,3,1,1,1,3,3,1,1,3,3,5,1,1,2,4,2,1,3,1,5,3,1,2,3,3]
[5,3,2,3,1,4,2,2,1,3,1,2,5,2,4]
[3,3,1,2,2,3,1,4,3,3,1,1,2,2,2,1,3,3,1,3,1,3,3,1,1,1,1,1,3,1,3,4,2,1,3,1,2,5]
[4,2,3,2,1,3,3,4,4,3,1,2,3,3,4,1,3,6,1,1,7]
[3,1.5,1,4,1,1,1,3,2,3,1,3,2,2,2,5,1,2,1,1,2,3,2,2,3,1,3,1,3,4]
[1,1,2,4,3,1,1,1,4,1,1,2,1,2,2,5,2,2,1,4,2,2,3,2,1,5,2,4]
[3,1,2,1,3,1,1,3,3,6,1,2,1,1,3,1,2,5,1,2,1,7,2,1]
[2,3,6,2,1,5,1,2,1,4,2,3,1,3,1,2,1,4,4,3,1,2,2,3,1]
[1,2,3,1,2,3,1,4,3,1,2,2,3,1,3,3,4,1,5,3,2,2,1,3,2,2,2,1,1,4,1,3,2,1,1,1,4,1,2,2,3,1,4,4]
[2,4,1,1,4,7,3,8,1,7,1,1,4,4,1,4,1,4,7]
[7,2,2,2,2,3,5,1,5,3,2,2,1]

Text 2.

a) POS sentence-wise:

[DET,N,A,CON,P,N,N,A,V,P,N,CON,P,DET,N,N,P,N,A,V,V]
[A,N,N,P,N,V,CON,N,CON,N,ADV,P,DET,N,N,N,P,PRO,V]
[P,N,P,N,A,N,V,CON,ADV,PRO,P,N,CON,N,P,PRO,V]
[P,N,N,ADV,P,N,V,P,P,N,A,A,V,ADV,DET,N,RA,V,CON,P,N,DET,N,A,N,P,N,P,N,V]
[ADV,V,CON,P,N,N,P,N,N,P,N,N,P,N,CON,N,V]
[N,CON,N,V,CON,P,N,N,P,N,N,CON,N,ADV,V]
[N,P,N,A,N,N,PRO,V,CON,N,N,A,RA,P,N,A,V,ADV,P,N,N,A,P,N,V]
[ADV,IF,N,P,N,A,PRO,P,N,A,A,N,CON,N,ADV,P,N,N,A,V,P,N,A,PRO,P,N,ADV,V]
[N,N,N,CON,N,N,N,P,N,CON,N,N,CON,N,P,N,V,CON,N,N,P,N,PRO,P,N,A,P,N,CON,N,P,PRO,V]
[ADV,N,N,N,PN,N,CON,N,A,N,P,N,N,A,CON,A,V,CON,N,A,V,A,N,CON,N,N,P,PRO,V]
[N,N,N,ADV,ADV,V,N,N,A,P,N,AUX,P,N,V,CON,N,N]
[PRO,CON,N,N,N,A,A,V,A,N,V,N,CON,N,P,PRO,P,N,A,V,N,CON,N,ADV,P,DET,N,V,CON,A,N,P,N,N,N,A,V]
[DET,N,A,N,A,A,A,P,N,A,P,N,N,CON,N,A,N,A,P,N,N,V,CON,ADV,N,RA,P,N,PRO,V]
[P,N,A,N,P,N,A,N,CON,N,CON,N,A,N,A,V]
[N,A,N,A,P,N,N,CON,P,N,N,N,N,A,P,N,N,V,P,A,N,P,DET,N,V]
[N,V,P,N,P,N,P,N,N,N,DET,N,RA,P,N,N,V,CON,P,N,N,PRO,P,N,A,CON,N,N,N,ADV,ADV,A,V]

[N,N,N,N,V,CON,N,A,A,N,N,PRO,RA,ADV,V,CON,ADV,ADV,P,N,N,V]
[N,A,P,N,A,DET,N,ADV,V,CON,N,N,A,CON,ADV,N,A,RA,V,CON,PRO,ADV,A,N,V,CON,IF,N,A,RA,V,N,A,PRO,RA,P,N,A,N,CON,V]
[ADV,A,A,N,ADV,P,N,CON,N,V,CON,ADV,N,N,P,N,N,P,N,RA,V]
[ADV,N,A,ADV,P,N,A,V,CON,V,CON,N,N,A,V,CON,PRO,RA,A,V]
[ADV,N,PRO,PRO,V,CON,N,N,CON,N,N,N,A,P,N,P,N,N,V,CON,PRO,N,CON,N,V,N,P,N,A,A,P,N,N,A,V]
[P,N,N,N,N,A,N,P,N,N,PRO,P,N,CON,N,V]
[P,A,N,V,CON,N,P,N,A,N,A,CON,P,N,CON,CON,P,N,PRO,P,N,V]
[PRO,V,CON,N,N,A,P,N,N,N,N,N,A,V]
[PRO,P,N,V,CON,N,N,A,RA,P,N,V,CON,P,N,N,P,N,CON,N,A,CON,P,N,N,A,V,N,CON,N,N,CON,N,A,N,RA,P,A,N,N,N,V]
[ADV,P,N,N,CON,N,N,CON,N,A,N,N,N,N,N,CON,V,CON,A,V,CON,N,CON,N,P,N,V,CON,N,V,CON,AUX,N,N,RA,V]
[N,ADV,P,N,N,A,P,N,N,A,P,N,CON,AUX,N,V,CON,PRO,PRO,RA,P,N,V,CON,N,ADV,CON,P,N,N,A,CON,N,PRO,P,N,N,V]
[P,A,N,N,A,V,P,N,A,A,A,N,A,RA,V,CON,P,N,P,N,P,N,A,P,N,P,N,A,V]
[ADV,N,CON,V,CON,PRO,V,N,PRO,V,CON,N,N,A,N,P,N,RA,PRO,V]
[N,V,P,N,P,N,ADV,P,N,N,N,DET,N,RA,P,N,N,V,CON,N,N,N,P,N,RA,V]

b) Word length sentence-wise:

[1,3,3,1,3,8,2,2,3,3,4,1,1,1,3,2,1,6,2,4,4]
[3,3,7,1,8,7,1,3,1,2,1,1,1,3,4,3,3,1,6]
[1,5,1,2,3,3,3,1,2,2,1,7,1,4,1,2,7]
[4,4,4,3,1,8,4,1,1,2,3,3,6,2,1,2,1,3,1,1,1,1,2,1,5,3,2,4,8,7]
[2,6,1,2,3,8,2,10,2,3,2,3,2,2,1,3,3]
[8,1,3,4,1,2,2,2,1,2,3,1,2,1,4]
[3,1,2,4,3,7,1,3,1,2,1,4,1,1,5,2,5,4,1,3,5,2,1,2,3]
[4,2,8,1,5,3,1,3,3,4,3,3,1,2,4,1,1,6,2,4,1,3,5,1,1,6,2,5]
[5,4,5,1,2,4,2,5,2,1,2,2,2,2,1,3,1,1,2,1,5,4,1,1,5,3,1,4,1,2,1,1,6]
[7,2,4,5,4,3,3,1,3,3,3,1,2,7,2,1,5,1,1,3,3,5,2,3,1,2,1,1,1,6]
[2,4,2,1,7,3,3,3,4,4,5,2,1,2,4,1,2,3]
[2,1,3,3,1,2,3,4,2,2,1,1,2,1,3,1,1,4,2,3,5,1,1,3,3,1,1,3,3,1,1,1,1,2,2,3,8,4]
[1,2,1,3,3,4,3,1,4,3,3,8,2,1,3,3,4,4,2,3,3,1,1,3,1,1,1,4,1,4]
[3,7,2,4,1,3,5,3,1,3,1,2,5,2,6,1]
[1,5,4,2,1,1,5,1,1,2,3,2,2,4,1,1,2,1,1,5,3,1,1,3,1]

[3,3,1,5,1,5,1,5,2,3,1,1,1,1,1,4,4,1,3,3,2,1,1,4,5,13,3,2,1,3,,1]
[5,2,2,3,1,2,2,3,4,1,4,1,1,4,4,1,2,2,1,1,5,7]
[2,4,1,2,2,1,2,3,4,1,3,2,4,1,3,5,5,1,5,1,1,3,1,1,2,1,2,3,5,1,6,3,2,1,1,1,2,3,2,1,5]
[4,3,3,4,2,1,5,1,2,5,1,3,3,4,3,3,2,1,5,1,8]
[2,8,4,2,1,5,3,3,1,5,1,3,3,5,1,1,1,1,3,3]
[5,4,1,1,1,1,3,3,1,4,4,2,3,1,3,2,3,7,7,2,2,2,1,2,3,1,1,3,3,7,1,4,4,4,3]
[1,2,1,10,3,8,2,1,2,2,1,3,3,1,2,2]
[1,2,4,1,1,3,3,3,3,4,4,1,1,3,1,1,1,4,1,4,3,5]
[1,3,1,3,2,4,1,3,2,6,4,2,4,7]
[1,2,2,1,1,4,3,2,1,4,5,6,1,2,2,2,1,2,2,1,2,3,3,1,3,3,4,3,3,3,1,1,2,4,1,2,2,1,1,3,1,2,2,2,4]
[4,1,4,3,1,5,4,1,2,2,1,3,3,6,5,1,1,1,3,1,1,4,1,4,1,5,2,1,6,4,1,3,2,4,1,4]
[4,2,4,3,2,4,1,2,3,3,1,2,1,5,2,3,1,1,2,1,1,3,5,1,4,2,1,3,2,5,3,1,1,2,1,3,2,3]
[1,2,4,3,3,3,1,4,3,2,3,2,2,1,3,1,2,5,1,5,1,2,3,1,4,1,5,3,4]
[3,4,1,5,1,2,4,3,2,5,1,3,2,4,3,1,2,1,2,7]
[3,3,1,5,1,5,3,1,5,2,3,1,1,1,1,1,4,4,1,2,4,4,1,3,1,4]

Text 3.

a) POS sentence-wise:

[N,N,N,N,V,CON,N,A,A,ADV,N,N,PRO,RA,V,CON,ADV,P,N,N,V,CON,N,ADV,A,N,RA,CON,V]
[N,V,N,A,N,CON,N,RA,P,N,A,P,N,N,N,N,CON,ADV,V]
[N,AUX,N,P,N,A,P,N,N,RA,P,N,V,CON,ADV,N,N,V,CON,ADV,P,N,V,N,RA,ADV,V]
[N,A,ADV,P,N,A,V,CON,N,A,P,N,N,A,P,N,A,V]
[ADV,N,N,N,A,N,A,N,A,RA,V,CON,N,N,CON,N,N,N,A,RA,V]
[N,N,CON,N,N,CON,N,RA,P,A,N,N,N,V,CON,V,A,N,CON,P,PRO,V]
[N,CON,N,P,N,N,P,N,A,AUX,P,N,A,CON,P,N,PRO,V,N,A,RA,V,CON,ADV,ADV,V]
[N,P,N,P,N,N,P,N,N,V,P,ADV,V,CON,ADV,P,N,N,N,RA,V,CON,AUX,N,PRO,RA,P,N,V]
[N,N,N,N,N,P,N,CON,P,N,V,P,N,N,CON,N,A,N,N,DET,N,N,CON,N,V]
[P,N,N,N,A,N,P,DET,N,P,N,A,CON,N,A,CON,A,P,A,N,N,CON,N,V]
[N,P,N,N,A,CON,P,PRO,N,P,N,P,N,CON,N,N,CON,P,N,P,N,N,A,V,CON,P,PRO,V]
[N,N,N,A,N,P,N,N,P,N,N,N,N,P,N,A,A,V]
[N,N,N,N,N,P,N,N,N,N,N,N,A,N,P,N,N,RA,P,N,N,A,DET,N,V]
[A,N,N,N,A,P,N,A,N,RA,P,N,N,P,N,N,N,A,P,DET,N,A,V]

[N,N,N,N,A,N,RA,V]
[P,N,N,N,N,ADV,N,N,N,N,N,P,N,N,V,P,N,N,P,N,A,V]
[N,N,N,N,N,A,N,RA,CON,P,N,N,A,P,N,P,N,ADV,PRO,P,N,V,V]
[N,P,N,P,N,P,N,V,CON,ADV,N,A,RA,A,V,CON,ADV,CON,AUX,PRO,V]
[DET,N,N,N,P,N,A,N,A,P,N,N,A,N,CON,N,A,N,ADV,P,N,A,N,A,P,DET,N,P,N,N,V]
[A,N,A,N,V,P,N,P,N,A,CON,N,A,P,N,N,P,N,A,A,N,V]
[N,N,N,V,CON,N,PRO,N,P,N,A,P,N,A,N,N,ADV,A,N,N,A,N,N,RA,V]
[N,N,P,N,N,P,N,A,P,N,N,CON,N,A,P,DET,N,V]
[N,N,N,P,N,N,N,PRO,P,N,P,N,N,A,N,V]
[N,A,N,V,N,N,N,N,RA,P,N,N,N,A,N,V
[N,N,N,N,A,N,N,CON,P,N,A,P,N,V,P,N,N,P,N,N,A,N,V,CON,ADV,N,A,CON,A,N,A,DET,N,N,RA,V]
[ADV,N,A,A,N,A,V,CON,A,N,A,N,P,DET,N,RA,P,N,N,N,V]
[N,N,ADV,V,CON,P,N,PRO,N,A,N,N,RA,N,N,CON,A,N,V,ADV,V,CON,N,N,A,N,RA,AUX,P,N,V]
[N,N,N,N,V,N,N,N,P,N,A,N,A,V,CON,N,N,N,P,N,P,N,A,PRO,P,N,N,CON,N,N,P,N,CON,A,V]
[N,N,N,A,P,N,N,DET,N,P,N,N,P,N,N,A,P,N,V]
[P,A,N,N,N,N,N,N,CON,N,N,N,P,N,V]

b) Word length sentence-wise:

[5,2,2,3,1,2,2,3,4,4,1,4,1,1,4,1,2,1,1,5,6,2,4,2,2,2,1,1,5]
[3,3,2,2,1,1,4,1,4,6,2,1,3,3,4,2,1,2,5]
[3,4,5,3,4,2,1,2,4,1,2,1,3,1,3,2,4,1,1,3,1,4,7,4,1,3,3,6]
[8,4,2,P,5,3,V,2,3,4,P,4,2,4,P,4,2,1]
[3,4,2,2,3,2,2,4,4,1,3,1,1,4,1,2,3,4,3,1,3]
[3,1,1,2,4,1,4,1,3,1,2,2,2,4,1,3,1,4,1,1,2,3]
[4,1,4,2,2,3,1,1,5,5,1,4,3,1,4,4,3,6,4,2,1,3,2,3,2,4]
[4,1,2,1,2,4,1,4,3,4,12,5,1,21,2,3,4,1,2,1,4,1,1,1,1,2,3]
[4,4,3,4,3,1,6,1,1,2,5,1,3,3,1,3,1,4,2,1,2,2,1,3,1]
[1,3,5,2,1,2,1,1,2,4,4,4,1,5,3,1,3,2,1,2,1,1,3,2]
[2,3,2,3,5,1,1,1,1,1,2,1,4,1,3,6,1,1,3,1,3,11,2,4,1,1,3,5]
[2,2,2,3,1,1,4,4,1,2,2,3,2,1,1,2,3,3]
[3,4,2,3,3,1,2,3,3,7,4,2,3,1,2,2,1,4,3,3,3,1,2,3]
[2,3,3,3,2,1,4,3,3,1,1,2,3,3,1,2,4,5,1,1,3,3,3]

[6,1,4,3,4,5,1,3]
[1,3,5,3,1,3,2,3,4,4,5,1,1,3,1,1,2,4,1,3,3,3,1]
[2,2,3,2,3,3,4,1,1,1,4,6,3,4,6,1,3,3,2,1,3,7,4]
[3,3,3,1,2,1,3,2,2,4,4,2,1,4,1,1,2,1,5,2,5]
[1,2,1,2,3,3,2,4,3,3,5,4,,3,1,2,2,2,3,4,3,1,4,4,3,1,3,1,1,5,3]
[1,2,2,3,1,4,5,3,7,3,1,6,3,1,5,3,1,2,2,3,1,5]
[3,2,4,3,1,2,1,2,1,5,3,1,3,4,2,4,3,2,2,2,3,3,2,1,2]
[4,4,1,3,6,1,3,4,3,5,4,1,3,5,1,1,2,4]
[9,3,8,1,3,3,3,1,1,3,4,3,2,2,3,3]
[3,3,3,6,5,2,4,5,1,1,1,3,2,3,4,4]
[3,3,2,2,3,2,3,1,3,3,4,1,1,4,2,2,3,1,5,3,3,1,1,1,4,2,3,1,1,3,3,1,2,6,1,5]
[3,1,2,3,4,2,4,1,1,3,3,4,1,1,2,1,1,2,2,5,6]
[3,1,3,1,1,1,3,1,3,3,1,3,1,4,3,1,2,3,4,2,2,1,5,2,3,3,1,1,5,3,4,5]
[4,3,4,3,1,6,4,3,1,3,1,2,1,3,1,2,2,3,4,3,1,1,2,1,3,2,4,1,3,3,1,4,1,1,3,1]
[3,4,2,4,1,4,4,1,2,1,2,2,1,2,1,2,3,2,4]
[1,2,3,5,4,2,3,2,1,1,1,2,1,4,4]

Text 4.

a) POS sentence-wise:

[N,V,N,A,N,P,N,N,P,N,A,V]
[P,N,N,N,N,P,N,P,N,A,N,P,N,N,PRO,RA,N,A,CON,N,N,P,PRO,V]
[N,N,N,A,N,N,N,N,A,N,N,N,PRO,RA,A,V]
[N,N,A,P,N,A,N,PRO,P,N,N,N,V]
[PRO,N,N,N,N,N,N,RA,P,A,N,N,P,N,N,V]
[N,N,N,ADV,V,CON,P,A,N,N,A,P,N,N,P,N,A,A,N,P,N,A,N,RA,V]
[P,N,N,N,N,N,V,CON,DET,N,A,P,N,N,P,N,A,N,A,CON,A,A,P,ADV,N,A,V]
[P,DET,N,ADV,P,N,DET,N,P,N,CON,N,N,A,N,CON,P,N,CON,N,N,V]
[N,A,N,P,N,A,P,N,N,A,P,N,N,P,N,V,CON,P,N,A,N,P,A,N,N,V]
[N,N,V,CON,N,PRO,N,A,A,P,N,A,P,A,N,ADV,N,P,A,N,A,RA,V,CON,CON,V]
[DET,N,P,N,A,CON,N,N,N,ADV,RA,V,V]
[PRO,P,N,V,CON,N,V,CON,P,N,N,N,V,CON,N,A,RA,V]
[P,N,A,N,A,P,N,N,P,N,DET,N,A,N,N,V,CON,A,N,A,V]
[N,N,N,P,N,P,N,N,N,P,N,N,V,N,N,CON,N,ADV,A,P,PRO,V]
[P,N,N,N,N,N,N,N,CON,N,PRO,RA,P,N,CON,N,A,V]

[ADV,P,DET,N,N,P,N,V,N,V,CON,N,A,N,RA,P,DET,N,V]
[ADV,DET,N,N,A,P,N,CON,N,PRO,V]
[ADV,N,N,DET,N,V,CON,ADV,CON,P,N,N,P,N,PRO,V,CON,ADV,ADV,P,N,PRO,V,CON,N,A,RA,P,N,PRO,V]
[N,A,V,CON,A,N,P,N,N,V,CON,N,N,P,N,N,RA,V,CON,AUX,ADV,V]
[ADV,N,A,N,CON,N,P,N,A,CON,ADV,P,N,N,P,N,N,A,V,CON,N,N,PRO,RA,V]
[N,A,N,CON,N,P,N,N,CON,N,P,N,N,A,V,CON,N,CON,N,N,P,N,A,N,P,N,N,P,N,A,V]
[N,N,N,P,N,CON,N,N,ADV,P,A,N,N,P,N,CON,N,ADV,N,A,CON,N,N,CON,ADV,N,A,A,N,N,CON,N,V]
[V,N,N,P,A,N,A,P,N,N,A,P,N,ADV,V,CON,N,A,V]
[N,N,P,N,N,N,A,V,ADV,N,A,P,N,CON,N,A,CON,P,N,V]
[P,AMN,N,A,P,N,N,CON,N,A,N,A,CON,,P,N,N,CON,N,V,P,N,N,A,N,N,A,N,V]
[N,DET,N,P,A,CON,A,CON,P,N,P,N,A,PRO,ADV,P,N,A,V]
[ADV,N,ADV,A,N,RA,CON,V,CON,V,N,CON,N,A,N,N,RA,P,N,A,P,N,V]
[ADV,N,P,DET.N,V,CON,N,N,A,CON,A,P,N,N,CON,N,CON,N,A,N,A,V]
[N,N,N,P,N,N,N,N,N,N,CON,N,V]
[ADV,ADV,P,N,P,N,P,N,N,N,A,N,N,A,P,N,N,N,P,N,A,RA,V]

b) Word length sentence-wise:

[2,1,2,5,3,1,1,3,3,2,3,2]
[1,3,5,7,2,2,2,1,3,2,3,1,2,2,1,1,3,2,1,2,3,1,1,3]
[5,2,3,5,2,2,5,5,2,2,5,2,1,1,3,V]
[5,2,3,2,3,3,3,1,1,3,3,9,1]
[1,2,3,10,3,3,9,1,1,4,3,3,1,2,3,2]
[4,3,2,2,4,1,1,1,2,3,3,1,2,4,1,2,3,2,1,1,4,3,3,1,3]
[1,3,5,3,4,2,5,1,1,3,3,1,2,2,1,2,3,2,2,1,3,3,1,3,1,3,3]
[1,1,3,3,1,3,1,4,3,3,1,2,3,4,3,1,1,2,1,2,3,8]
[3,4,4,5,3,5,1,2,6,2,5,4,3,1,4,3,1,24,2,1,1,3,4,4,5]
[2,3,5,1,4,1,1,33,2,3,4,1,3,4,2,4,2,2,5,2,1,2,1,1,4]
[1,3,3,4,4,1,2,2,2,4,1,5,3]
1,1,2,1,1,4,6,1,2,4,2,4,5,1,2,4,1,7]
1,2,3,4,1,2,4,1,2,1,2,1,2,2,2,1,1,1,2,4]
[5,4,3,1,4,1,3,4,2,1,1,4,1,3,3,1,1,2,2,1,1,4]
[[4,1,3,3,2,1,1,2,1,31,1,4,4,1,3,2,3]
[3,4,1,2,3,3,2,4,4,3,1,3,2,2,1,1,1,2,6]

[3,1,4,3,8,1,3,1,3,2,2]
[3,2,2,1,2,1,1,3,3,1,1,3,1,3,1,6,2,3,1,1,2,1,6,1,4,3,1,1,1,1,5]
[8,4,5,1,2,3,1,2,1,6,1,4,3,1,3,4,1,6,1,3,3,6]
[4,2,3,5,1,3,1,2,3,1,1,1,4,2,1,3,2,2,1]
[2,2,3,1,3,1,3,2,1,3,1,2,3,3,1,1,2,1,2,2,1,3,2,2,1,3,3,1,2,3,1]
[3,4,3,1,6,1,3,2,3,4,3,4,2,1,3,1,2,3,2,4,1,2,3,1,1,3,6,1,2,1,1,6,2]
[6,4,3,2,2,2,3,3,2,1,5,1,7,3,4,1,3,2,7]
[4,4,2,3,5,5,3,2,3,3,3,1,2,1,3,3,1,1,3,5]
[1,2,2,1,3,3,3,3,1,2,1,2,2,1,3,3,2,1,3,4,1,1,4,4,3,2,3,8,4]
[3,1,2,4,3,1,4,1,3,3,1,3,3,1,2,1,4,3,5]
[2,4,2,2,2,1,1,5,1,4,4,1,1,2,2,1,1,1,5,2,1,5,4]
[3,2,1,1,2,1,1,3,1,2,1,4,1,3,5,1,4,1,6,3,3,3,5]
[2,5,2,1,4,1,3,4,4,4,1,4,5]
[4,5,2,2,1,2,1,2,10,3,1,5,1,4,2,2,2,2,1,3,2,1,4]

Text 5.

a) POS sentence-wise:

[N,V,N,CON,N,P,N,V,CON,N,P,N,N,V]
[N,N,N,CON,N,N,N,N,A,PRO,RA,P,N,N,V]
[ADV,N,P,P,ADV,PRO,V,CON,N,A,P,N,A,V,ADV,A,V,CON,P,PRO,P,N,N,V,CON,P,N,A,CON,A,PRO,V]
[P,DET,N,N,N,V,CON,N,A,A,N,RA,P,N,V,CON,ADV,A,N,RA,P,N,N,P,N,V]
[N,N,P,N,N,P,PRO,V,CON,N,N,N,A,ADV,P,A,N,P,N,P,N,N,P,N,P,N,PRO,V]
[N,N,V,CON,DET,N,CON,N,PRO,ADV,ADV,P,N,N,CON,N,V,N,A,V,ADV,N,P,PRO,V]
[N,P,N,PRO,V,ADV,A,N,RA,P,N,P,N,N,P,N,V]
[N,N,A,N,P,N,P,N,P,N,A,N,P,N,A,P,N,ADV,V,CON,N,A,N,P,DET,N,V]
[N,CON,N,P,DET,N,V,CON,P,N,N,P,N,A,N,A,N,P,N,P,N,A,V]
[N,A,P,N,V]
[N,N,DET,N,N,PRO,P,N,RA,V]
[N,P,N,A,P,N,CON,P,PRO,AUX,P,N,A,V]
[N,CON,P,N,A,A,P,N,V,AUX,P,N,N,A,P,PRO,V]
[DET,N,P,N,A,N,N,N,CON,N,PRO,P,N,A,PRO,CON,A,N,P,N,P,N,N,A,N,A.V]
[ADV,P,N,P,PRO,CON,DET,N,A,N,P,N,ADV,ADV,V]
[DET,N,AUX,A,N,CON,N,CON,P,N,N,N,P,DET,N,V,V]
[A,N,N,N,AUX,N,DET,N,RA,V,CON,N,PRO,RA,P,N,N,V]

[IF,PRO,N,N,RA,V,PRO,AUX,P,N,A,P,N,N,V]
[N,N,N,A,DET,N,RA,P,A,N,A,P,N,N,V]
[P,N,N,N,N,A,N,A,P,N,N,A,N,CON,N,P,N,P,A,N,A,A,A,A,CON,A,V]
[ADV,N,A,A,P,N,A,N,A,N,A,RA,P,N,P,N,A,P,N,V]
[N,ADV,N,N,PRO,RA,P,N,N,CON,P,N,A,N,V,V]
[N,N,V,N,P,N,CON,CON,V,N,A,A,CON,A,N,RA,V]
[A,N,A,V,N,ADV,N,CON,N,P,N,V]
[ADV,ADV,N,P,N,N,A,P,N,CON,N,RA,P,N,CON,N,N,P,N,A,V]
[N,N,N,P,N,A,PRO,V]
[N,N,N,N,CON,N,ADV,P,N,P,PRO,V]
[A,V,P,A,N,A,N,CON,A,N,N,A,N,P,N,V]
[A,N,A,N,P,N,DET,N,V,N,PRO,RA,P,A,N,A,N,N,N,P,N,N,V]
[N,V,N,P,N,A,CON,ADV,P,N,N,V]

b) Word length sentence-wise:

[3,4,4,1,4,1,5,2,1,3,3,2,4,4]
[5,5,3,1,3,3,1,4,2,1,1,1,5,1,5]
[3,2,1,1,2,1,1,1,4,3,1,2,2,1,1,3,1,1,1,1,1,2,2,8,1,1,4,2,1,3,2,7]
[4,1,3,4,4,5,1,3,3,2,4,1,1,3,6,1,7,3,3,1,3,1,3,1,2,2]
[3,2,1,3,4,2,1,4,1,3,5,2,3,2,2,2,1,1,3,4,2,4,1,3,1,2,1,4]
[2,1,4,1,1,4,1,4,1,2,2,1,2,3,1,3,3,2,3,3,2,8,3,1,8]
[4,P,4,1,6,2,1,2,1,3,4,1,2,1,3,2,4]
[3,2,2,3,1,9,1,3,1,1,5,3,1,3,3,3,2,44,1,2,5,3,1,1,1,5]
[2,1,3,2,1,1,4,1,1,2,3,3,2,2,2,3,3,1,3,6,4,2,8]
[4,3,3,3,8]
[4,2,1,2,1,2,1,7,1,8]
[3,1,2,3,2,3,1,2,1,2,1,3,1,5]
[3,1,3,4,5,3,1,2,7,2,2,2,7,5,1,2,4]
[1,2,3,3,1,2,1,9,1,1,2,1,4,4,1,1,1,N,3,3,1,4,2,5,4,2,1]
[4,1,4,1,2,3,1,2,4,3,3,2,3,6,5]
[1,4,4,3,3,1,4,1,1,2,3,4,2,1,4,6,2]
[1,2,2,2,4,2,1,2,1,5,1,1,1,1,2,4,9,4]
[2,2,3,2,1,4,2,2,3,2,3,1,4,2,5]
[4,9,3,3,1,2,1,3,3,4,2,1,2,3,5]
[5,2,2,1,4,5,5,4,1,2,2,3,2,1,2,1,4,1,3,2,3,4,3,3,1,3,3]

[3,4,3,3,1,4,3,3,2,3,4,1,1,3,1,5,7,1,2,3]
[3,6,3,3,1,1,4,5,3,1,1,6,3,3,5,5]
[3,4,5,4,4,3,1,3,1,2,3,3,4,1,4,2,1,5]
[1,4,2,1,4,3,2,1,4,3,2,3]
[5,2,4,1,3,2,3,1,2,1,1,1,1,5,1,4,5,1,2,2,3]
[5,2,3,1,4,3,1,6]
[4,2,1,3,1,4,1,4,2,1,1,4]
[3,1,5,3,3,2,1,5,2,3,3,3,1,4,4]
[1,2,4,4,1,2,1,2,1,3,1,1,1,1,4,4,2,1,2,1,2,2,4]
[2,1,3,1,4,3,1,5,1,2,3,6]

J.W.v. Goethe, *Der Erlkönig*, Lemmas replaced by their frequencies

1,2,4,1,1,1,9,2,
13,6,13,6,13,5,
13,3,13,2,1,4,13,3
13,2,13,1,13,2,13,1

24,4,2,1,13,4,1,13,1,
1,9,13,13,3,4,
13,3,6,1,9,1,
24,4,13,6,2,1,

13.1,5,1,2,6,24,
1,3,1,1,24,6,13,
2,1,1,6,4,13,1,
24,1,3,2,1,1,

24,9,24,9,9,1,13,4,
2,3,24,1,1,
6,2,1,2,24,5,
4,1,1,1,13,2,

1,1,2,13,6,24,2,
24,3,1,13,1,3,
24,3,1,13,1,1,
9,1,9,1,9,1,13,

4,9,4,9,9,3,13,4,1,
3,3,4,1,1,
24,4,24,4,24,3,13,1,

13,1,13,1,1,4,1,

24,1,13,24,1,13,3,1,
9,6,13,4,1,4,1,24,1,
4,9,4,9,1,2,13,24,
3,3,24,2,1,1,

13,9,1,13,2,1,
13,2,4,3,13,1,5,
1,13,1,6,1,9,1,
4,13,3,13,5,6,1

Author Index

Subject Index

The RAM-Verlag Publishing House edits since 2001 also the journal ***Glottometrics* –** up to now 32 issues – containing articles treating similar themes. The abstracts can be found in http://www.ram-verlag.eu/journals-e-journals/glottometrics/.

The contents of the last issue (32, 2015) is as follows:

Herausgeber – Editors of Glottometrics