Glottometrics 34
2016

RAM-Verlag

ISSN 2625-8226

Glottometrics

Glottometrics ist eine unregelmäßig erscheinende Zeitdchrift (2-3 Ausgaben pro Jahr) für die quantitative Erforschung von Sprache und Text.
Beiträge in Deutsch oder Englisch sollten an einen der Herausgeber in einem gängigen Textverarbeitungssystem (vorrangig WORD) geschickt werden.
Glottometrics kann aus dem **Internet** heruntergeladen, auf **CD-ROM** (in PDF Format) oder in **Buchform** bestellt werden.

Glottometrics is a scientific journal for the quantitative research on language and text published at irregular intervals (2-3 times a year).
Contributions in English or German written with a common text processing system (preferably WORD) should be sent to one of the editors.
Glottometrics can be downloaded from the **Internet**, obtained on **CD-ROM** (in PDF) or in form of **printed copies.**

Herausgeber – Editors

External academic peers for Glottometrics

Bestellungen der CD-ROM oder der gedruckten Form sind zu richten an
Orders for CD-ROM or printed copies to RAM-Verlag RAM-Verlag@t-online.de
Herunterladen/ Downloading: https://www.ram-verlag.eu/journals-e-journals/glottometrics/

Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Glottometrics. 34 (2016), Lüdenscheid: RAM-Verlag, 2016. Erscheint unregelmäßig.
Diese elektronische Ressource ist im Internet(Open Access) unter der Adresse
https://www.ram-verlag.eu/journals-e-journals/glottometrics/ verfügbar.
Bibliographische Deskription nach 34 (2016) **ISSN 2625-8226**

Contents

Glottometrics 34, 2016, 1-8

Syntactic Text Characterisation Using Linguistic S-Motifs

Reinhard Köhler, Sven Naumann, Trier[1]

Abstract. R-motifs, formed from categorical instead of quantitative sequences, are used to characterise texts with respect to parts of speech as a syntactic feature. The attempt to classify the end-of-year speeches of the Italian presidents by means of not more than some parameters of the of the thus formed motifs fails. Instead, it could be shown that Ord's criteria and two other characteristics of the empirical frequency distributions of the motifs reveal a continuous development over time which follows the Piotrowski-Altmann Law.

Keywords: *motif, S-motif, text characterisation, syntactic properties, parts of speech, Piotrowski-Altmann Law*

Introduction

Motifs as a new kind of unit were introduced into linguistics in Köhler (2006 and 2008a) to enable quantitative researchers to study the syntagmatic structure of texts on a scalable level of granularity. The unit was previously inspired by a contribution by Boroda (1982) to musicology, where a unit, which could be considered as an equivalent to the linguistic word was not available, and rhythmic structures using Zipf's law could not be conducted. Boroda defined his F-motif as the sequence of notes with increasing duration. Linguistic motifs are defined in a more general way as

> *the longest continuous sequence of equal or increasing values representing a quantitative property of a linguistic unit.*

All kinds of linguistic units with their quantitative properties can be chosen as basic units to form motifs from the numerical values, such as *morphs, words or syntactic constructions* with the individual measures of their properties such as *length, frequency, polysemy, polytextuality, age*, etc. The corresponding motifs are called L-, F-, P-, T-, and A-motifs respectively. In Beliankou, Köhler, Naumann (2013), another variant of motifs was introduced to open up the option to conduct corresponding investigations also on the basis of categorical properties. This variant is defined as

> *An R-motif is an uninterrupted sequence of unrepeated elements.*

An example of the segmentation of a text fragment (represented as a sequence of argumentative relations) into R-motifs is the following:

["elaboration"], ["elaboration", "concession"], ["elaboration", "evidence", "list", "preparation", "evaluation", "concession"], ["evidence", "elaboration", "evaluation].

According to this definition, a current motif ends as soon as the first repetition of an item occurs. This repeated token does not belong to the current motif but becomes the first element of the successor motif.

[1] University of Trier, Germany. Address correspondence to: koehler@uni-trier.de

The present paper is an experiment to characterise texts by means of as simple as possible methods using motifs on the basis of syntactic properties, called S-motifs. The only syntactic feature we will use in this study is the sequence of part-of-speech categories, and the only kind of motif is the simple R-motif of PoS tags, i.e. S-motifs, on the first level of granularity.

Data and method

The experiment is conducted on a text corpus[2] consisting of 63 end-of-year speeches of the Italian presidents from 1949 to 2011, where the word sequences were replaced by the corresponding sequences of PoS tags. As an example, the first sentence of the first end-of-year speech of the first Italian president Einaudi (1949) becomes

[PREP N] [PREP A N V] [PREP A N PRON V] [N V PREP DET A] [N DET] [N PREP] [N PRON , CONG PREP] [N PREP A] [N, AVV PREP A] [N PRON] [PRON V PREP DET N] [PREP A N CONG] [PREP A N].

The S-motifs are formed, as usual, regardless of punctuation and syntactic analysis of the sentence structure. For each text, the frequencies of the resulting motifs were determined and corresponding frequency spectra were formed. The Waring distribution (cf. e.g., Wimmer and Altmann 1999)

$$P(x) = \frac{b}{n}\frac{n^{(x)}}{(b+n+1)^{(x)}}, \quad x = 0,1,2,\ldots, b > 0, n \geq 0, \tag{1}$$

where

$$n^{(x)} = n(n-1)(n-2)\ldots(n+x-1), x \in \Re, n \in N$$

was fitted to the obtained spectra. Fitting was successful with good and very good goodness-of-fit values in all cases except the first three speeches delivered by president Einaudi, whose texts were very short and yielded not enough data classes for the Chi-square test. His next two speeches became somewhat longer and could be fitted by the Waring distribution also with good results. The empirical parameters of the short speeches could be evaluated nonetheless, of course. Table 1 is an example of the fitting results of the Waring distribution to the data.

Table 1
Fitting the Waring distribution to the spectrum of the PoS motifs
in the first speech by president Gronchi (1955)

x_i	f_i	NPi
1	80	80.00
2	8	7.99
3	2	2.16
4	1	0.84
5	1	1.01
Parameters	b = 2.8489	n = 0.4273
P(χ^2) = 0.992	DF = 1	$ORD_{b,S}$ = <0.334; 2.4603>

[2] Many thanks to Arjuna Tuzzi, University of Padua, for these data.

Results and Observations

Our first attempt was to find out whether the statistical characteristics of the spectra could be used to classify the texts and whether the resulting classes would correspond to the presidents in a similar way as we did with some success in Köhler, Naumann (2008c). In that earlier study we tried to separate text sorts whereas our present approach concentrated on the authors of the texts (all of which belong to an identical text sort and language) and uses as little as possible statistical effort. Classification algorithms, however, failed to produce any useful grouping of the texts. So, we tried other methods which have proved to be successful in many cases.

It is common practice in quantitative linguistics to compare and differentiate data sets (e.g., representing a property of texts) by means of Ord's criteria *I* and *S* (cf. Ord, 1972) which characterise the position of a given empirical frequency distribution on a two-dimensional plane. Often, similar texts form clusters and can thus be separated from other data sets. This method was, to some extent, successful also with the Italian presidents, however not fully satisfying. The diagrams in Fig. 1a-1d show visualisations of Ord's criteria. Here, the data sets were beforehand divided into four groups of the apparently mutually most similar ones. The differences between the groups as well as the similarities within the groups can be seen from the differently scaled axes. The lines in the diagrams do not belong to the graphs of the criteria; they have been inserted to show which of the points belong to which speech (cf. Table).

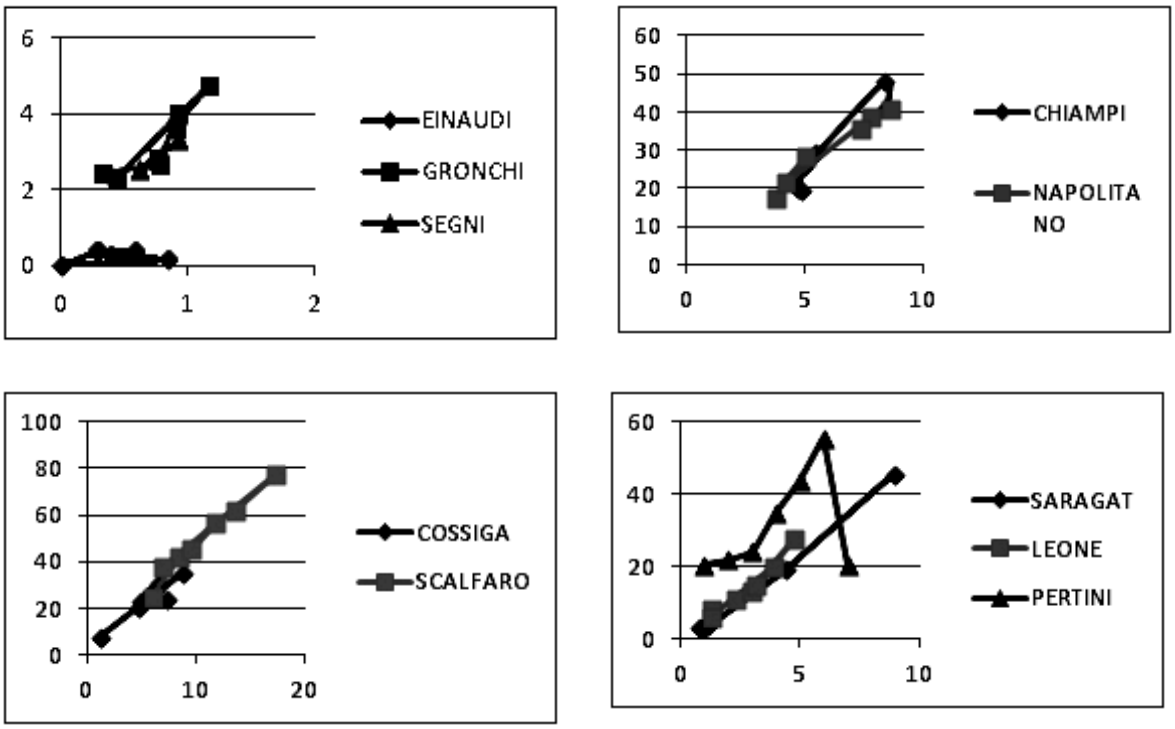


Fig. 1 a-d: Four groups of presidential speeches according to Ord's criteria

Another observation can be made by looking on the interrelations between the years of the speeches and the characteristics in Table 2. The values of Ord's *I* and *S*, the theoretically calculated frequencies of the first frequency classes of the individual distributions NP_1, and the first moments (i.e. the average values) on the other hand, seem to systematically increase with the year of the speeches. Figs. 2-5 show graphically the corresponding values (marks) together with the theoretical functions (solid lines). The functions are the results of fitting the logistic function (1), which is known in linguistics as the Piotrowski-Altmann Law, to the

data sets. Usually, this function is applied in line with Altmann's derivation of the law from a differential equation under a spatial or social interpretation: A population is hypothetically divided into two groups of which one is the growing one which already uses a new linguistic form and the other one is the decreasing one which still resists the innovation under study – or the other way round in case of the dynamics of a loss of a property. Our present hypothesis is based on a similar interpretation but we cannot determine or assume a fixed population, of which we could also determine the proportion of the 'infected' members. The reason is that the population changes, while the growth process is active. New presidents replace the old ones, new members of the writer group replace older ones and the group sizes may change, too. Nonetheless, we assume that the same mechanism can serve as a (more abstract) model of the dynamics under consideration and that the application of the Piotrowski-Altmann Law is justified.

The axes in the diagrams 2-5 represent the dependent variable on the vertical (*y*-axis) and time in steps of 10 years on the horizontal one (*x*-axis).

(2) $$pt = \frac{c}{1 + ae^{-bt}}$$

As can be seen, each of the characteristics increases indeed with time. Ord's criteria follow clearly the usual saturation function with its first, slowly growing part, the middle part with a rapid increase followed by a slow-down part, and finally the typical tail with a slower and slower approximation to the saturation value. Both data sets display not only increasing values but also remarkably increasing variances.

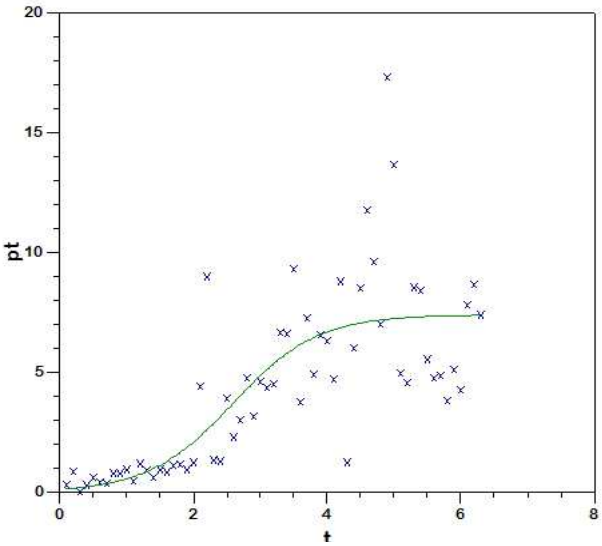


Fig.2: The Piotrowski-Altmann function as fitted to Ord's *I.* The estimated parameter values are a = 59.8341, b = 1.5790, and c = 7.4016. The value of the determination coefficient is R^2 = 0.5704.

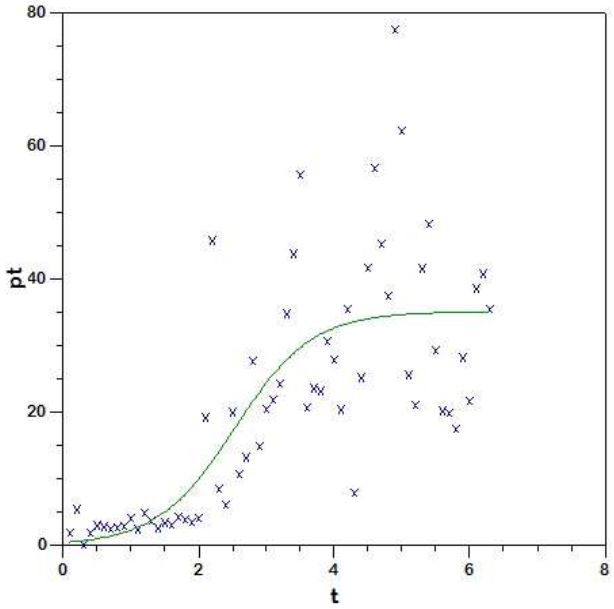


Fig. 3: The Piotrowski-Altmann function as fitted to Ord's *S*. The estimated parameter values are a = 83.21147, b = 1.7477, and g = 35.1037. The value of the determination coefficient is $R^2 = 0.5642$.

A slightly different picture is obtained with the NP_1 values. Here, the data begin shortly after the turning point of the function and the variances appear to scatter more smoothly around the theoretical function.

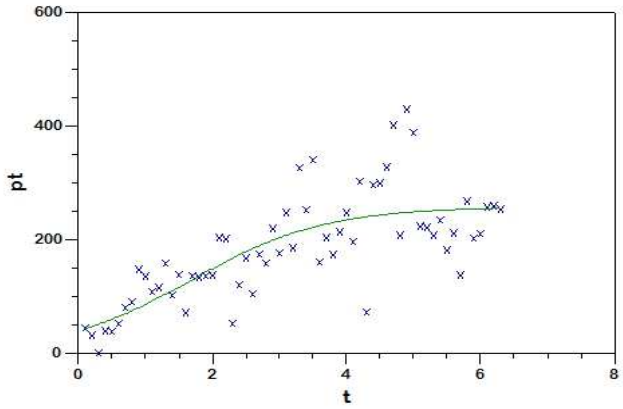


Fig. 4: The Piotrowski-Altmann function as fitted to the NP_1 values. The estimated parameter values are a = 5.4713, b = 1.0069, and c = 257.7609. The value of the determination coefficient is $R^2 = 0.5625$.

Finally, function (1) was fitted to the values of the mean values of the distribution. These data points start apparently in the middle of the last part of the function, indicating that a preliminary end of the dynamics of this characteristic is reached.

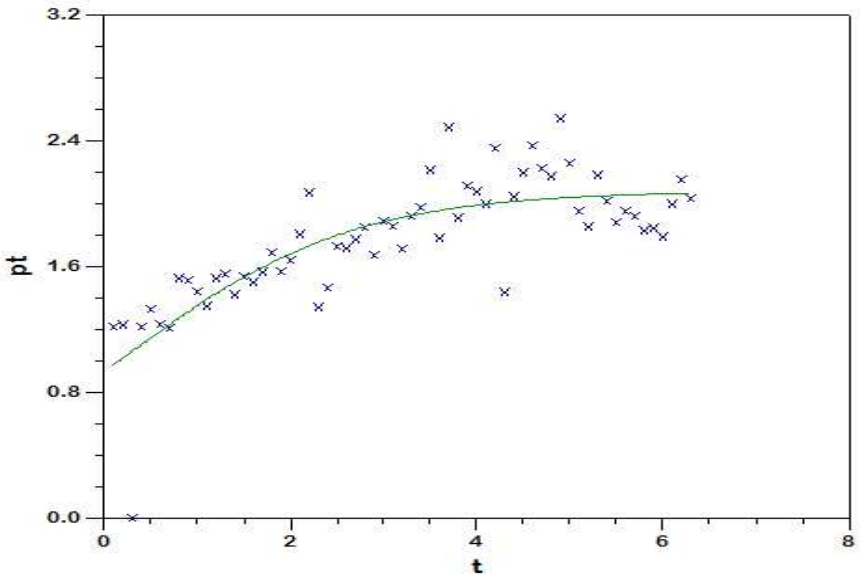


Fig. 4: The Piotrowski-Altmann function as fitted to the m'_1 values. The estimated parameter values are a = 1.2302, b = 0.8232, and c = 2.0809. The value of the determination coefficient is $R^2 = 0.6425$.

Table 2
Some characteristics of the distributions of S-motif properties in the end-of-year speeches of the Italian presidents from 1949 to 2011

Speech (No., president, and year)	**Ord's *I***	**Ord's *S***	NP_1	m'_1
01_EINAUDI_1949	0.3004	19.919	44.264	1.2157
02_EINAUDI_1950	0.8412	10.6637	31.2871	1.2286
03_EINAUDI_1951	0	13.1279	0	0
04_EINAUDI_1952	0.2826	27.5773	39	1.2174
05_EINAUDI_1953	0.5919	14.86	38	1.3261
06_EINAUDI_1954	0.3838	20.4171	52.4353	1.2295
07_GRONCHI_1955	0.334	21.8033	80	1.2065
08_GRONCHI_1956	0.7765	24.2225	89.9546	1.5246
09_GRONCHI_1957	0.7698	34.7295	147	1.5126
10_GRONCHI_1958	0.9291	43.6885	135.3203	1.4419
11_GRONCHI_1959	0.4371	55.6218	107.5658	1.3478
12_GRONCHI_1960	1.1681	20.5795	115	1.5235
13_GRONCHI_1961	0.9058	23.4673	157.7379	1.5509
14_SEGNI_1962	0.616	23.1308	102	1.4211
15_SEGNI_1963	0.9299	30.6083	138.5191	1.5372
16_SARAGAT_1964	0.8333	27.8422	71	1.5
17_SARAGAT_1965	1.1085	20.3133	135.6461	1.5628
18_SARAGAT_1966	1.1608	35.362	133.2776	1.6872
19_SARAGAT_1967	0.9292	7.8262	135.9891	1.5691
20_SARAGAT_1968	1.2197	25.0792	137.1554	1.6406
21_SARAGAT_1969	4.3929	41.5978	203	1.8038
22_SARAGAT_1970	8.9703	56.6416	201	2.0676
23_LEONE_1971	1.3239	45.204	51.6365	1.3387
24_LEONE_1972	1.2681	37.3597	119.7813	1.4641
25_LEONE_1973	3.8951	77.4624	167	1.7269

26_LEONE_1974	2.2726	62.141	104	1.7153
27_LEONE_1975	2.9906	25.4764	174	1.7729
28_LEONE_1976	4.7532	21.0493	158.1755	1.8458
29_LEONE_1977	3.1418	41.483	219	1.6737
30_PERTINI_1978	4.6161	48.1764	176	1.8893
31_PERTINI_1979	4.3485	29.1559	247.1431	1.8563
32_PERTINI_1980	4.4867	20.0814	185	1.7113
33_PERTINI_1981	6.6479	19.7365	326	1.9203
34_PERTINI_1982	6.6061	17.3633	252.5785	1.9762
35_PERTINI_1983	9.2876	28.0937	340	2.2124
36_PERTINI_1984	3.7549	21.6271	159.4812	1.7797
37_COSSIGA_1985	7.2663	38.5069	203	2.4837
38_COSSIGA_1986	4.9073	40.6738	173	1.9071
39_COSSIGA_1987	6.5376	35.4568	213	2.1115
40_COSSIGA_1988	6.3157	19.919	247	2.0754
41_COSSIGA_1989	4.6998	10.6637	196.3834	1.9966
42_COSSIGA_1990	8.7823	13.1279	302	2.3501
43_COSSIGA_1991	1.2167	27.5773	72	1.4348
44_SCALFARO_1992	5.985	14.86	296	2.0433
45_SCALFARO_1993	8.4981	20.4171	299	2.1962
46_SCALFARO_1994	11.7492	21.8033	327	2.3692
47_SCALFARO_1995	9.6087	24.2225	401	2.2255
48_SCALFARO_1996	6.9864	34.7295	207	2.1722
49_SCALFARO_1997	17.2939	43.6885	429	2.5405
50_SCALFARO_1998	13.6472	55.6218	388	2.2546
51_CIAMPI_1999	4.9383	20.5795	223	1.9539
52_CIAMPI_2000	4.546	23.4673	221	1.8515
53_CIAMPI_2001	8.5188	23.1308	207	2.1806
54_CIAMPI_2002	8.4017	30.6083	234	2.0186
55_CIAMPI_2003	5.5179	27.8422	181	1.8795
56_CIAMPI_2004	4.7649	20.3133	211	1.9536
57_CIAMPI_2005	4.8657	35.362	137.1348	1.9196
58_NAPOLITANO_2006	3.8127	7.8262	267	1.8291
59_NAPOLITANO_2007	5.0991	25.0792	202.4369	1.8448
60_NAPOLITANO_2008	4.2379	41.5978	210	1.7875
61_NAPOLITANO_2009	7.7988	56.6416	257	1.9971
62_NAPOLITANO_2010	8.6459	45.204	259	2.1528
63_NAPOLITANO_2011	7.3771	37.3597	253	2.0313

Discussion and Perspectives

We were not able to classify the end-of-year speeches of the Italian presidents using the simple criteria on which we tried several algorithms. On the one hand, some of the texts showed clear particularities, which can be expected to differentiate them from others. An example is Einaudi with his modest, uniquely short and straight speeches. On the other hand, some displayed overlapping property behaviour such as Cossiga and Scalfaro with long and complicated texts, again others such as Ciampi and Napolitano could hardly be differentiated. Moreover, the dynamics shown in Figs. 2-5 display a more or less continuous development of the values. Even Einaudi's speeches become longer after some time. This situation can be interpreted in several ways. One of them is the assumption that the (ghost) writer teams of the presidents developed their style continuously and not abruptly with changes of the presidents, possibly because not the complete teams were replaced when a president replaced the previous one. Another interpretation assume that the grammatical-stylistic differences in the texts

as a general development within the text sort "end-of-year speeches" more than in personal style.

We will have to wait for more investigations of texts on the basic of S-motifs and their characteristics on various text sorts and several languages, including the dynamic aspect, before a preliminary solution to this problem is possible. What we can say now is that the applied, very simple measure of syntactic complexity on the basis of part-of-speech R-motifs, here called S-motifs, seems to yield encouraging outcomes. In follow-up studies, we will (1) contribute to more investigations based on S-motifs, (2) find out how motifs of higher order, e.g. length motifs, frequency motifs etc. of S-motifs, length motifs of length motifs of S-motifs etc. – which enable us to create characteristics with a larger scope but with lower resolution – can provide us with more useful information for classification purposes in a similar way as in Köhler, Naumann, (2008c), and (3) conduct experiments with other simple measures of syntactic complexity.

References

Beliankou, Andrei; Köhler, Reinhard; Naumann, Sven (2013). Quantitative Properties of Argumentation Motifs. In: Obradović, Ivan; Kelih, Emmerich; Köhler, Reinhard [eds.], *Methods and Applications of Quantitative Linguistics. Selected papers of the VIIIth International Conference on Quantitative Linguistics (QUALICO) in Belgrade, Serbia, April 16-19, 2012: 33-43.* Belgrade.

Boroda, Moisei (1982). Häufigkeitsstrukturen musikalischer Texte. In: Orlov, Jurij K.; Boroda, Moisei G.; Nadarejšvili, Isabela Š. [eds.], *Sprache, Text, Kunst. Quantitative Analysen*: *231-262*. Bochum: Brockmeyer.

Köhler, Reinhard (2006). *The frequency distribution of the lengths of length sequences.* In: Genzor, J.; Bucková, M. (eds.), *Favete linguis. Studies in honour of Victor Krupa: 145-152.* Bratislava: Slovak Academic Press.

Köhler, Reinhard (2008a). Word length in text. A study in the syntagmatic dimension. In: Mislovičová, Sibyla (ed.), *Jazyk a jazykoveda v prohybe*: *416-421*. Bratislava: VEDA vydavatel'stvo SAV.

Köhler, Reinhard (2008b). Sequences of linguistic quantities. Report on a new unit of investtigation. *Glottotheory 1(1), 115-119.*

Köhler, Reinhard; Altmann, Gabriel (1996). "Language Forces" and synergetic modelling of language phenomena. In: Schmift, P. (ed.), *Glottometrika 15: 63-76*. Trier: WVT.

Köhler, Reinhard; Naumann, Sven (2008c). *Quantitative text analysis using L-, F- and T-segments*. In: Preisach, Burkhardt; Schmidt-Thieme, Decker (eds.), *Data Analysis, Machine Learning and Applications*: 637-646. Berlin, Heidelberg: Springer,

Köhler, Reinhard; Naumann, Sven (2009). *A contribution to quantitative studies on the sentence level*. In: Köhler, Reinhard (ed.), *Issues in Quantitative Linguistics: 34-57*. Lüdenscheid: RAM-Verlag.

Köhler, Reinhard; Naumann, Sven (2010). A syntagmatic approach to automatic text clas sification. Statistical properties of F- and L-motifs as text characteristics. In: Grzybek, Peter; Kelih, Emmerich; Mačutek, Ján (eds.), *Text and Language: 81-89.* Wien: Prae sens.

Ord, J. Keith (1972). *Families of frequency distributions*. London: Griffin.

Wimmer, Gejza; Altmann, Gabriel (1999), *Thesaurus of univariate discrete probability distributions*. Essen: Stamm.

Glottometrics 34, 2016, 9-13

The Relationship between English Synonyms and Compounds

Hanna Gnatchuk[1]

Abstract: The present study is concerned with the detection of a relationship between the number of a word's synonyms and the number of compounds a word can have. In such a way, we deal with two variables in the research – the number of synonyms and the number of compounds. The links have been captured by means of a very simple exponential equation. As a result, the connection between the analyzed variables has been positively confirmed.

Key words: synonyms, compounds, quantitative linguistics.

1. Introduction: the problems of synonyms in quantitative linguistics

In order to get a clear idea on the possible problems of synonymy in quantitative linguistics, it would be relevant at first to give a definition of a term "synonym". According to Oguy (2003), synonyms are "sinnverwandte Morpheme bzw. Wörter mit unterschiedlicher lautlicher Form und gleicher oder ähnlicher Bedeutung, die denselben Begriff oder sehr ähnliche Begriffe ausdrücken" (2003: 90-91). On the whole, one distinguishes in linguistics three types of synonyms: *semantic (ideographic), stylistic and contextual*. The semantic synonyms are the words which differ in the shades of their meanings to a certain extent: *English*: big – large; happy – lucky; to say – to tell; *German*: sprechen – reden, Gehalt – Lohn – Gage; Geruch – Gestank – Duft; *French*: laisser – quitter, petit – minime; *Ukrainian*: сміливий – відважний – мужній, шлях – дорога, страх – жах; *Russian*: кроткий – покорный – незлобивый. As far as stylistic synonyms are concerned, they deal with the usage of words in different styles: *Germ*: Gesicht (neutral) – Antlitz (poetic) – Fresse – Fratze; *French*: visage – museau; *English*: stop – cease; *Russian*: спать – почивать, лицо – лик – рожа – морда; *Ukrainian*: їсти – жерти, працювати – трудитися – ішачити. The contextual (or occasional) synonyms are the words, „die in synonymischen Wörterbüchern kodifiziert sind, unterscheiden sich okkasionelle Synonyme, die üblicherweise im Text entstehen" (Oguy, 2003: 93). The contextual synonyms are the words which have the same meaning only in a certain context. But they are not synonyms without this context.

Levickij (2012) considers that one of the most important problems of synonymy in quantitative linguistics is: *what do the number of a word's synonyms depend upon?* Here a researcher puts forward the following simple hypothesis: the number of synonyms depend upon the number of a word's meanings. Although he emphasizes the importance of studying the connection between semantic, stylistic, morphological aspects of a word and the number of a word's synonyms. In such a way, A. Vengrynovych (2003, 2005, 2005) conducted a quantitative study of synonyms on the basis of three dictionaries of synonyms (*Duden, Görner/Kempcke, Bulitta*). As a result, 35834 nouns were found in the dictionary by Duden;

[1] Hanna Gnatchuk, Universität Trier, Computational Linguistics and Digital Humanities, Universitätsring, 15. Email: agnatchuk@gmail.com or s2hagnat@uni-trier.de

10950 nouns – in the dictionary by Bulitta; 17292 – in the dictionary by Görner/Kempcke. A great number of synonyms were found in the dictionary by Bulitta. The researcher explains it by the fact that the dictionary by Bulitta belongs to the so-called cumulative types of dictionaries. The obtained data have been processed by means of chi-squared test and coefficient of contingency. In such a way we shall regard the results of the research done by Vengrynovych by considering semantic, stylistic, morphological and frequency aspects of words, on the one hand, and their number of synonyms, on the other hand:

1) Semantics and the number of synonyms

Vengrynovych (2005) classified the synonyms from three dictionaries into 23 semantic subclasses, namely a semantic subclass of a) abstract notions; b) the names of organizations and departments; c) substance; d) feeling; e) form and structure, etc. The maximum number of synonyms of a word was 7. In such a way, a semantic subclass "names of organizations" has the following frequencies: 1 synonym = 810 words; 2 synonyms = 87 words; 3 synonyms = 33; 4 synonyms = 21 words; 5 synonyms = 12 words; 6 synonyms = 14 words; 7 and more = 49. The total number of words is 1026.

The received data have been treated with the help of the contingency coefficient and chi-squared test which helped the researcher to find statistically significant connections between certain semantic subclasses and the number of synonyms:

Apartment = 1 synonym
People = 2 synonyms
Plants = 2 synonyms
Abstract notions = 7 or more synonyms, etc

In such a way, Vengrynovych stated that abstract nouns have a higher propensity to have more synonyms than concrete ones.

2) Stylistics and the number of synonyms

Another aspect is a stylistic class of a word (neutral, colloquial, dialectal, poetic words, etc). Here a researcher was interested in the questions whether the number of synonyms depends upon a word's stylistic class. A statistical analysis has shown that there is a statistically significant connection between the following classes of words: colloquial words (colloquial style) = 2 synonyms or 7 or more synonyms; emotional words = 1 synonym or 7 or more synonyms; dialect words = 2 synonyms; neutral words = 5 or 6 synonyms. Judging from the results, Vengrynovych made a conclusion that neutral words have more synonyms than colloquial, emotional, dialectal ones.

3) Morphology and the number of synonyms

In order to study the connection between a morphological aspect of a word and the number of synonyms, Vengrynovych divided all the synonyms into three morphological subclasses in reference to their gender – words with feminine, masculine and neutral gender. Moreover, the author distinguishes simple, derived and complex nouns. Having treated the data statistically, he has found that feminine nouns proved to have the highest number of synonyms. This can be explained by the fact that there are many abstract nouns with a feminine gender (i.e. –heit, -keit, etc). Simple and complex words proved to have a small number of synonyms in contrast to derived words (7 and more synonyms).

4) Frequency and the number of synonyms

Dealing with the frequency of nouns, Vengrynovych used a frequency dictionary by Ortmann. Unfortunately, not all nouns have been found in the dictionary under consideration. The same problem was in the research by Nemcova/Serdelova (2005). In such a way, Vengrynovyh distinguished 3 classes of frequencies: nouns with low, average and high frequencies. A statistical analysis has shown that low-frequent words have 1-2 synonyms, average and high frequent words – 7 or more synonyms.

In such a way, it is possible to make the following conclusions in regard to the number of synonyms based upon the results of the research conducted by Vengrynovych (2005, 2003):

- *High and average frequent words proved to have more synonyms than low-frequent ones;*
- *Stylistically neutral words turned out to have more synonyms than colloquial, dialect and poetic words;*
- *Feminine nouns proved to have more synonyms than neutral and masculine ones. This can be explained by the fact that the majority of feminine nouns belong to abstract nouns;*
- *Abstract nouns are inclined to have more synonyms than concrete ones;*
- *Simple and complex words have a small number of compounds in comparison with derived ones.*

2. Empirical testing a hypothesis on the relation between the number of synonyms and compounds

The objective of the study consists in confirming or refuting the connection between a word's synonyms and the number of word's compounds. The next step is to put forward zero (null) and alternative hypotheses which we shall deal with:

H_1: *If a word possesses a considerable number of synonyms, it is more inclined to build a great number of compounds*;

H_2: *The number of a word's synonyms do not influence the number of a word's compounds.*

Or, in other words, the more synonyms a word has, the more compounds it produces. Both synonyms and compounds are means of meaning specification hence they must be linked in some way.

The data of our study are represented by 744 English lexemes taken from *Webster's New Dictionary of Synonyms* (1973). In order to accept one of the above-mentioned hypotheses, we have taken the following steps:

- 744 English words have been taken (English words = the first element of compounds);
- We have calculated the number of synonyms for these 744 English words;
- The number of English compounds have also been computed for these 744 words.
- The results are illustrated in Table 1.

We propose to set up the differential equation:

$$dy/(y - a) = c*dx,$$

Table 1

The number of the first components of the compounds (N) and the number of synonyms the first components produce (C)

The number of synonyms x	**N**	**C**	**Observed values N/C = y**	**Computed**
1	11	21	1.50	4.24
2	47	165	3.51	4.32
3	80	400	5.00	4.44
4	135	819	6.06	4.61
5	155	854	5.50	4.88
6	122	765	6.27	5.28
7	82	575	7.01	5.89
8	49	231	4.70	6.82
9	34	291	8.6	8.21
10	21	236	11.23	10.31
11	4	67	16.75	18.27
12	4	105	26.25	25.51
a = 4.0901; b = 0.1006; c = 0.4123; R^2 = 0.9616				

We express the fact that the relative rate of change of *y* is constant. The parameter *a* in the denominator is a shift of the function. We obtain the result

$$y = a + b*\exp(c*x)$$

that is, a very simple exponential equation. The result of computation is as follows:

$$y = 4.0901 + 0.1006*\exp(0.4123x).$$

In such a way, the above-formulated hypothesis on English synonyms has been positively corroborated ($R^2 = 0.9616$). Nevertheless, it is necessary to investigate the behavior of synonyms and compounds in other languages with the aim of corroborating the above result and discovering a possible language laws.

References

Levickij, V. V. (2012). *Semasiologia* [*Semasiology*]. Vynnytsa: Nova Knyha (in Russian);

Oguy, O. D. (2003). *Lexikologie der Gegenwärtigen Deutschen Sprache*. Winnyts's „Nowa Knyha".

Vengrynovych, A. (2003). Zalezhnist mizh semantykoju imennuka i kil'kistu synonimiv u nimetskij movi [Dependence between the semantics of a noun and the number of sznonzms in German]. *Naukovyj visnuk Chernivetskoho universutetu: Germanska filologija, 168,* 23-31.

Vengrynovych, A. (2005). Synonimija ta morpholohichnuj status imennuka [Synonymy and a morphological status of a noun]. *Naukovyj visnuk Chernivetskoho universutetu: Germanska filologija, 231*, 65-73.

Vengrynovych, A., Levickij, V. V. (2005). Kolichestvenn'je parametr' sinonimii v nemetskom jaz'ke. In: Kvantitativnaja lingvistika: issledovanija i modeli (KLIM – 2005). Material' Vserossijskoj nauchnoj konferentsii (6-10 June, 2005), Novosibirsk, 228 – 232.

Webster's New Dictionary of Synonyms (1973). G & C. Merriam Company.

Glottometrics 34, 2016, 14-27

Quantitative Analysis of US Presidential Inaugural Addresses

Miroslav Kubát, Radek Čech[1]

Abstract. The research aims to investigate several features of inaugural addresses of the presidents of the United States. The goal of the paper is to observe the presidential speeches from a viewpoint of stylometry indices and to discover whether political and historical circumstances (wars, financial crisis, ideology, etc.) influence the style of inaugural addresses, analogically to findings presented by Čech (2014). Specifically, vocabulary richness, thematic concentration and text activity are computed. These three indices were chosen especially due to (a) their high efficiency of automatic text classification (genre analysis, authorship attribution, etc.), (b) their independence on text length and (c) simple linguistic interpretation. The combination of the three methods allows both to investigate the style of the particular presidential speeches in powerful linguistically comprehensive view and to observe the development trends of the specific genre of inaugural addresses during the more than 200 years long history. The corpus comprises inaugural addresses of all US presidents from George Washington to Barack Obama (57 texts in total).

***Keywords**: stylometry, presidential speeches, vocabulary richness, thematic concentration, activity*

1. Introduction

Political speeches are widely used in linguistic research, especially in discourse analysis (e.g. Lim, 2004; Carranza, 2008; Matić, 2012). Several quantitative analyses have dealt with this issue (e.g. Čech, 2014; Savoy, 2010; Tuzzi et al., 2010). It is not surprising therefore that addresses of the US presidents are frequently investigated because the American President can be ranked among the most powerful politicians of the contemporary world. In this study, we analyse all US presidential inaugural addresses. While most analyses deal with these data in terms of qualitative methods or content analysis, we focus on the issue from a viewpoint of stylometric indices of contemporary quantitative linguistics, particularly vocabulary richness, secondary thematic concentration, and text activity. These methods have proved to be an effective tool in political language research. Promising results were obtained by Čech (2014) who analysed an impact of ideology on a character of annual messages given by Czech and Czechoslovak presidents. Another related research was done by Tuzzi et al. (2010) who examined end-of-year speeches of Italian presidents.

The aim of this study is to analyse relationships between some characteristics of the style of US presidential speeches and certain pragmatic aspects which could have an impact on the addresses, specifically, historical development, ideology, financial crises, and wars. It is important to emphasise that this study is but a first insight into the issue and our approach is rather heuristic.

[1] University of Ostrava, Dept. of Czech Language. Czech Republic. Address correspondence to: cechradek@gmail.com

2. Language material

The inaugural address is a habitual part of the inauguration procedure. Except for constitutionally required presidential oath of office all parts of the inauguration procedure (including inaugural speech) are optional given by tradition. This is the reason why several presidents (particularly John Tyler, Millard Fillmore, Andrew Johnson, Chester A. Arthur, Calvin Coolidge) gave no address. In each of these cases, the incoming president substituted a president who had died.

This genre provides unique data for quantitative linguistic research because of homogeneity of the genre and its long tradition. In this study, 57 addresses were analysed. The list of all addresses with the results can be found in the appendix of the article. The data was collected by the American Presidency Project (Peters and Woolley, 2015).

3. Methodology

We use three methods to investigate some aspects of the style of US presidential addresses, mainly the vocabulary richness (*MATTR*), secondary thematic concentration (*STC*), and text activity (*Q*). These indices were chosen due to (a) high efficiency of automatic text classification (genre analysis, authorship attribution, etc.), (b) their independence on text length, and (c) simple linguistic interpretation. The vocabulary richness was computed by *MaWaTaTaRaD* software (Milička, 2013); the thematic concentration and the activity were computed by *QUITA - Quantitative Index Text Analyzer* (Kubát et al., 2014).

3.1 Moving Average Type-Token Ratio (MATTR)

The measurement of vocabulary richness is one of the oldest quantitative methods in stylometry, with more than seventy years long history (cf. Popescu et al., 2009). A large number of indices of vocabulary richness has been set up in linguistics; however, almost all of them evidence an undesirable dependence on the length of the text. To avoid this dependence in our analysis, we use the moving average type-token ratio (*MATTR*), proposed by Covington and McFall (2010), which was experimentally proved to be independent of the text size (see Kubát, 2014).

The *MATTR* is defined as follows. A text is divided into overlapped subtexts of the same length (so called "windows" with arbitrarily chosen size *L*; usually, the "window" moves forward one token at a time), next, the type-token ratio is computed for every subtext and, finally, the *MATTR* is defined as a mean of particular values. For example, in the following sequence of characters: *a, b, c, a, a, d, f,* text length is 7 tokens ($N = 7$) and we choose the window size of 3 tokens ($L = 3$). We get subsequent 5 windows: *a, b, c* | *b, c, a* | *c, a, a* | *a, a, d* | *a, d, f,* and compute *MATTR* of the sequence as follows:

$$MATTR(L) = \frac{\sum_{i=1}^{N-L} V_i}{L(N-L+1)} = \frac{3+3+2+2+3}{3(7-3+1)} = 0.87$$

L…arbitrarily chosen length of a window, $L < N$
N…text length in tokens
V_i…number of types in an individual window

3.2 Secondary Thematic Concentration

The secondary thematic concentration (*STC*) is a method which measures the degree of intensity with which the author focuses on a topic (or topics) of a given text (cf. Čech et al., 2015). Specifically, the *STC* is based on two text characteristics: 1) the frequency distribution of words and 2) the so called *h*-point (cf. Popescu, 2007). The *h*-point is defined as a point where the frequency equals rank (see formula 1); it separates in a fuzzy way the most productive synsemantics from autosemantics in a rank frequency distribution of words or lemmas (for more details cf. Popescu et al., 2009, p. 17ff). Specifically,

$$(1) \quad h = \begin{cases} r_i, & \text{if there is } r_i = f(r_i) \\ \dfrac{f(r_i)r_{i+1} - f(r_{i+1})r_i}{r_{i+1} - r_i + f(r_i) - f(r_{i+1})} & \text{if there is } r \neq f(r) \end{cases} ,$$

where r_i is the rank and $f(r_i)$ is the respective frequency of this rank; given that r_i is the highest number for which $r_i < f(r_i)$ and r_{i+1} is the lowest number for which $r_{i+1} > f(r_{i+1})$. Thus, if no rank is equal to the respective frequency, one computes the lower part of formula (1) consisting of neighbouring values. Having stated the *h*-point, all autosemantics occurring at lower ranks are considered as thematic words because they signalize the frequent repetition of the given autosemantics.[2] (Čech et al., 2015). The *h*-point is multiplied by two in the concept of the *STC*, on reasons presented in Čech et al. (2015). The thematic weight (*TW*) of each thematic word can be computed and, finally, the *STC* is obtained as the sum of these weights (*TW*), specifically

$$(2) \quad STC = \sum_{r'=1}^{2h} TW = \sum_{r'=1}^{2h} \frac{(2h - r')f(r')}{h(2h-1)f(1)} ,$$

where *r'* is the rank of autosemantic word above *h*-point and *h* is the *h*-point. For illustration, we present here the computation of the *STC* of the Lincoln's inaugural address (see the Text 20 in Appendix and Table 1).

Table 1
The rank-frequency distribution of Text 20. *h* = 9.

Token	**Rank**	**Average rank**	**Frequency**	**Token**	**Rank**	**Average rank**	**Frequency**
the	1	1	58	for	11	10	9
to	2	2	27	with	12	12.5	8
and	3	3	24	be	13	12.5	8
of	4	4	22	this	14	14.5	7
it	5	5	13	a	15	14.5	7
war	**6**	**6.5**	**12**	by	16	17.5	6
that	7	6.5	12	we	17	17.5	6
all	8	8	10	is	18	17.5	6
in	9	10	9	**god**	**19**	**17.5**	**6**
which	10	10	9				

[2] It should be mentioned that not all autosemantics need be considered to express the thematic properties of the text; for instance Popescu et al. (2009) use only nouns and their predicates of the first order, i.e. adjectives and verbs. In this paper, this approach is followed.

$$STC_{Text\ 20} = \sum_{r'=1}^{2h} \frac{(2h - r')f(r')}{h(2h-1)f(1)} = \frac{(2 \cdot 9 - 6.5)12}{9(2 \cdot 9 - 1)58} + \frac{(2 \cdot 9 - 17.5)6}{9(2 \cdot 9 - 1)58} = 0.0159$$

3.3 Activity

Each text focuses more intensively either on the action (plot) or on the description. For instance, travel books focus principally on description and, conversely, short stories concentrate on the plot. The concept of the activity and descriptiveness was introduced by Busemann (1925). Generally, the text activity is represented by verbs and the descriptiveness by adjectives. Index of activity *Q* is defined as a ratio of verbs *V* and the sum of verbs *V* and adjectives *A* in the text, see formula (3):

$$(3) \quad Q = \frac{V}{V + A}$$

For illustration, the activity *Q* of the Lincoln's inaugural address (Text 20 in Appendix) is

$$Q_{Text\ 20} = \frac{V}{V + A} = \frac{102}{102 + 36} = 0.74 \ ,$$

which expresses high activity of the text.

3.4 Statistical comparison

In this study, differences between results are tested by means of the *u*-test[3], see formula 4

$$(4) \quad u = \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad ,$$

where, $\bar{X}_1$, $\bar{X}_2$...arithmetic mean of results in each group,
S_1, S_2…standard deviation,
n_1, n_2…number of results in each group.

Since the threshold is 1.96, $u \geq 1.96$ means that the difference between two groups is statistically significant for the significance level $\alpha = 0.05$.

4. Results

4.1 Historical development

Firstly, we focus on the historical development of all US presidential inaugural addresses. The chronologically ranked resulting values are presented in Figure 1.

[3] In statistics, it is sometimes called *z*-test; here, we follow a convention used in the quantitative linguistics.

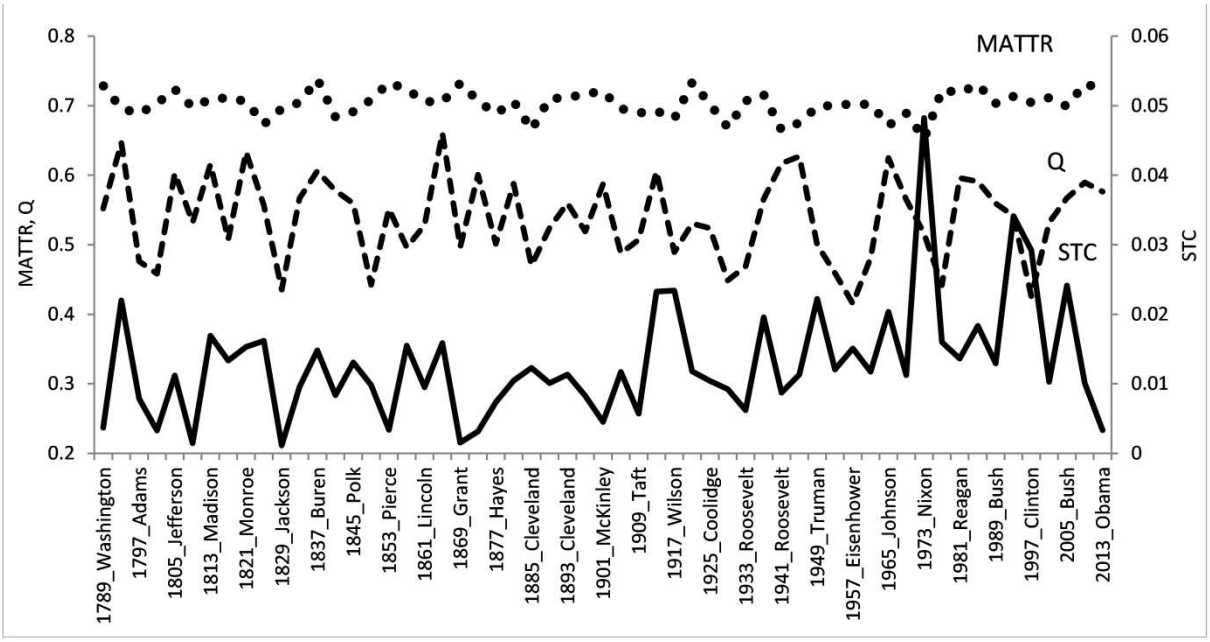


Figure 1. Chronologically ranked values of *MATTR*, *Q*, and *STC* of the US presidential inaugural addresses.

As can be seen in Figure 1, there is no tendency at first sight. The obtained values of the indices oscillate chronologically up and down without any obvious reason. The results seem to be a matter of individual style of each president rather than historical circumstances. Nevertheless, we try to find out whether the style of the addresses is influenced by some pragmatic causes, namely: political affiliation, war, and financial crisis.

4.2 Political affiliation

Throughout most of the American history, a two-party system dominated. Since 1852, every American president has been presented as a candidate either of Democratic or Republican political party. Before this date, the political affiliation of particular president was not so evident; consequently, we use only the addresses from 1852 for the analysis of the potential impact of political affiliation on the style. Theoretically, the political affiliation can influence political speeches because of different ideological basis (cf. Čech, 2014). Our aim is to discover whether inaugural addresses of democratic presidents differ from the republican ones. The resulting values are presented in Table 2.

Table 2
MATTR, *Q*, *STC* resulting values and statistical comparison of democrats and republicans

	democratic	republican	*u*
MATTR	0.70	0.70	0.13
Q	0.54	0.54	0.08
STC	0.015	0.012	0.82

The results in Table 2 show that there is no significant difference (at the significant level $\alpha = 0.05$). Surprisingly enough, the values of *MATTR* and *Q* display even no difference at all. Thus, we can state that political affiliation has no impact on the style of inaugural addresses in terms of the measured indicators. More detailed view of the issue is displayed in Figure 2 and 3 where the style of addresses is expressed as relation *MATTR-Q* and *MATTR-STC*.

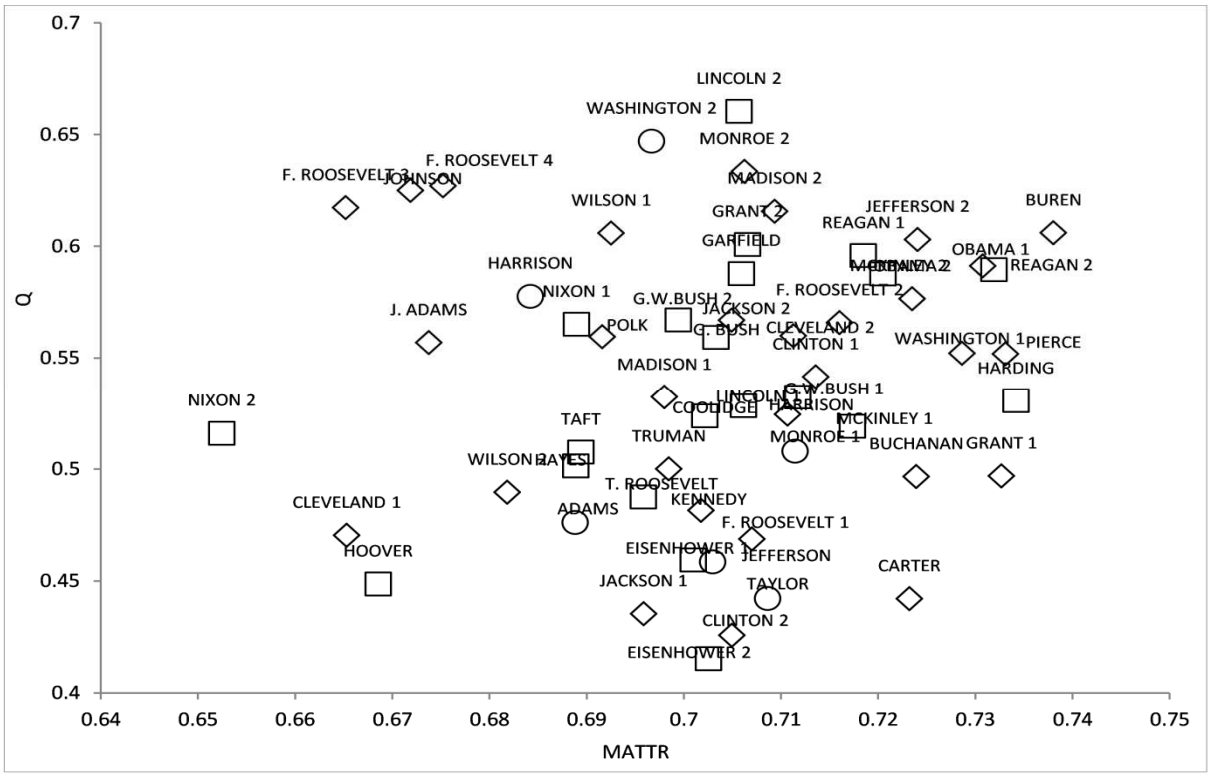


Figure 2. The relation between *MATTR* and *Q* in inaugural addresses;
square = republican, diamond = democratic, circle = others.

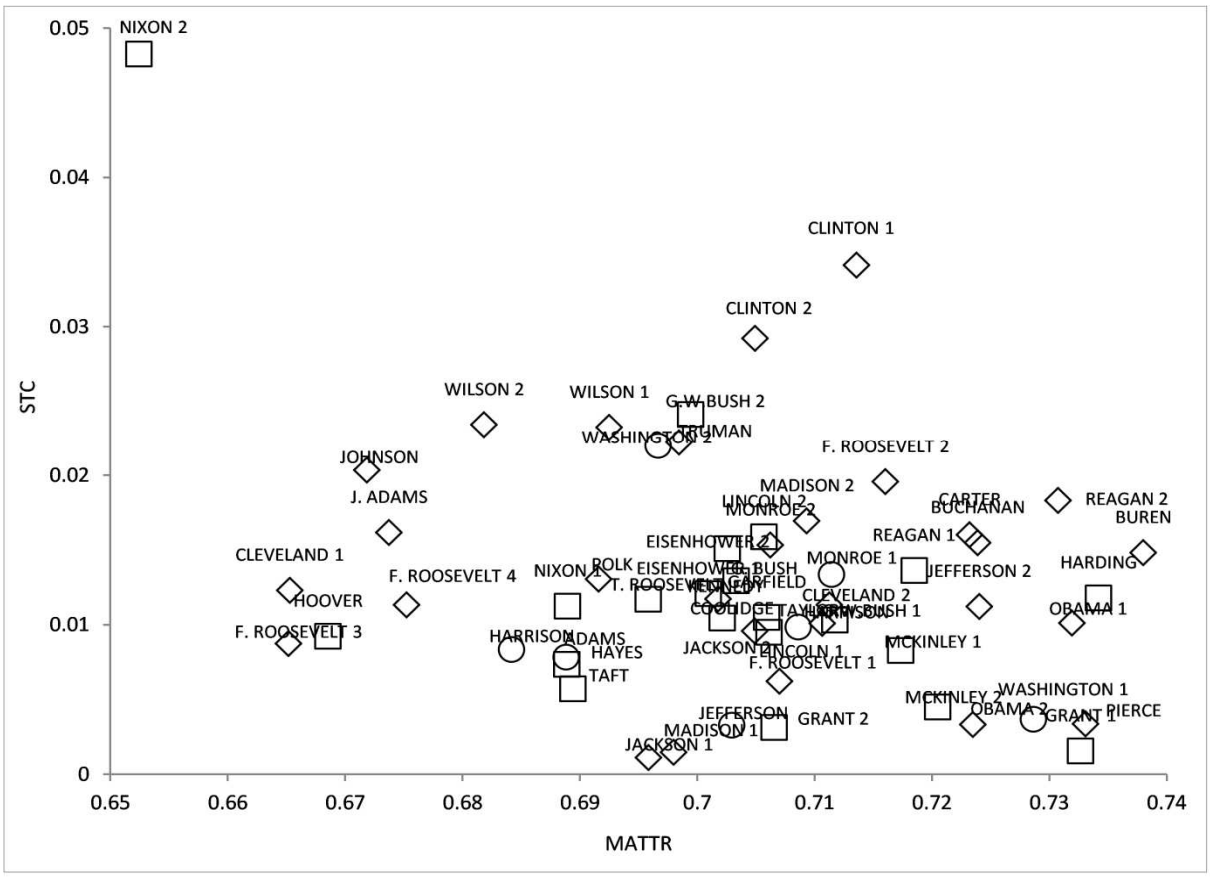


Figure 3. The relation between *MATTR* and *STC* in inaugural addresses;
square = republican, diamond = democratic, circle = others.

4.2 Wartime

A war affects a society in many ways, especially "big" ones such as the First and Second World War. For politicians, a war usually represents one of the most important topics in their political agenda and the wartime can be interpreted as an extraordinary era (in contrast to peacetime). This fact could be reflected by different style of wartime political speeches (in contrast to peacetime speeches).

However, the history of the USA, as of any other country, seems to be a series of various wars and it is difficult to decide which era can be assigned as the peacetime and which as the wartime. For example, let us consider the Cold War, the long era of strained and polarized relations between East and West. On the one hand, it was not a real war in fact; on the other hand, the cold war was one of the biggest wars in terms of number of arms, its impact to the particular societies, and danger of nuclear arms usage and so on. It is even hard to decide how long this war lasted.

Considering the aforementioned methodological problems, we decide to distinguish peacetime and wartime according to the US military expenditures (in percent of GDP). We choose 4% value as the border which seems to be suitable to distinguish the worst wars in US history (see Figure 4). Although this threshold is an arbitrary chosen value just for the purpose of this study, this method allows us to reasonably distinguish between wartime and peacetime.

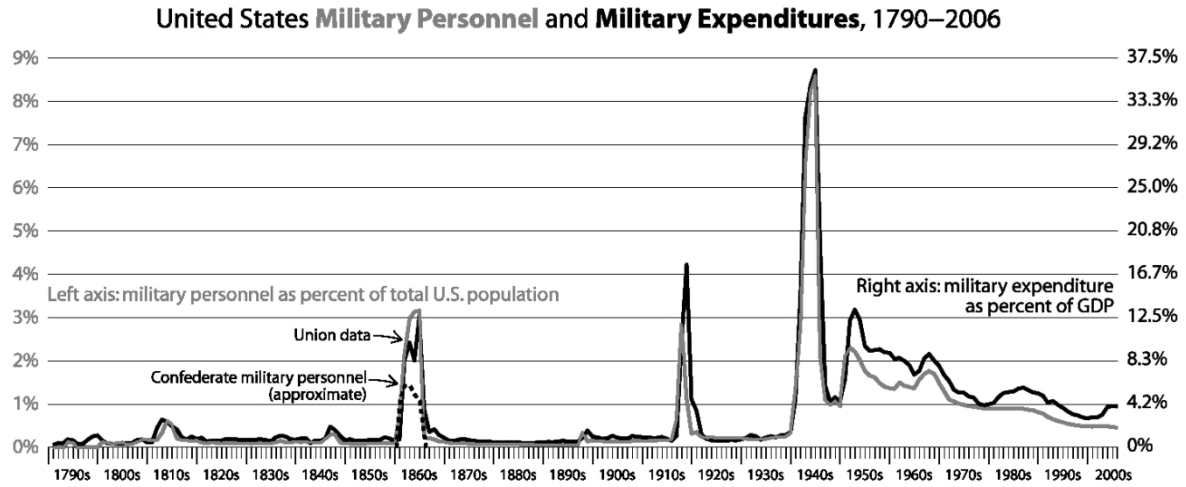


Figure 4. Military expenditures and military personnel in US history. Source: https://en.wikipedia.org/wiki/File:US_military_personnel_and_expenditures.png

As can be seen in Figure 4, only several wars are considered as wartime according to this criterion, namely the American Civil War, First World War, Second World War, Cold War, Korean War, Vietnam War, and partly ongoing War on Terror.

Table 3
MATTR, Q, STC resulting values and statistical comparison of wartime and peacetime.

	wartime	peacetime	*u*
MATTR	0.70	0.71	0.96
Q	0.55	0.53	0.94
STC	0.017	0.010	**2.98**

The results in Table 3 show that wartime and peacetime addresses significantly differ only in the case of the *STC*. This result is in accordance with our assumption (see above) and is probably caused by the fact that in wartime era the war is really dominant topic whereas in peaceful era president tends to talk about more topics. This statement can be supported by the findings of totalitarian language (Čech, 2014). The *MATTR* and *Q* rather reflect the style of speeches; the results reveal that it (at least in the case of observed characteristics) is not influenced by the wartime.

4.3 Financial crisis

Aside from war, recession is one of the worst eras for people. The financial crises often trigger strikes, social unrests, and sometime even wars. There are several options how to determine financial crises through the history. We decide to use the unemployment rate which influences significantly the standard of living and is directly caused by recession. Since we do not have data before 1890, we must analyse only the period after this year; moreover, the values of unemployment between 1890 and 1940 are only estimated. Nevertheless, from the Figure 5 is obvious that there are two extraordinary periods where the unemployment exceeded 10%, particularly 1894-1898 and 1931-1939. Five hundred banks closed, 15,000 businesses failed, and the unemployment hit 35% in New York and even 43% in Michigan in the first serious economic depression starting in 1883, just thirteen days before the inauguration of G. Cleveland. The second financial crisis known as "The Great Depression" was the longest, deepest and most widespread depression of 20th century. This crisis started after the stock market crash of October 29, 1929 (known as Black Tuesday). The effect on people was enormous: more than 5000 banks failed, unemployment rate exceeded 20%, and hundreds of thousands found themselves homeless. The resulting values of *MATTR*, *Q*, *STC* and statistical comparison are displayed in Table 3.

Figure 5. Unemployment rate in US history. Source: https://commons.wikimedia.org/wiki/File:US_Unemployment_1890-2009.gif

Table 3
MATTR, Q, STC resulting values and statistical comparison of normal and crisis

	normal	crisis	*u*
MATTR	0.70	0.71	**2.70**
Q	0.53	0.53	0.21
STC	0.016	0.011	1.35

As can be seen in Table 3, despite small difference in terms of *MATTR* resulting values (0.70 and 0.71), only vocabulary richness significantly distinguishes normal time and recession. Activity seems to be irrelevant in terms of crisis and despite some difference between thematic concentration results (0.016 and 0.011); the statistical test does not prove significant difference.

4.4 Thematic Words

The method of measurement of thematic concentration allows extraction of the so called thematic words, i.e. words which represent main topic(s) of text. The thematic words (*TW*) can be viewed as an alternative to keywords (cf. Čech et al. 2015). The advantage of *TW* lies in the fact that those words are based solely on the frequency structure distribution of the text; no reference corpus is needed for the analysis. The list of *TW* of all inaugural addresses is displayed in Table 4. The complete list of *TW* of each presidential speech can be found in the Appendix .

Table 4
Frequency list of thematic words of all inaugural addresses ($f \geq 2$).

#	**Word**	**f**		#	**Word**	**f**
1	have	46		23	freedom	5
2	government	29		24	other	5
3	people	28		25	citizens	4
4	has	22		26	war	4
5	been	20		27	law	3
6	world	13		28	let	3
7	who	11		29	what	3
8	country	11		30	nations	3
9	great	10		31	united	3
10	nation	10		32	own	3
11	states	10		33	had	3
12	shall	9		34	congress	2
13	more	9		35	power	2
14	America	7		36	free	2
15	peace	7		37	democracy	2
16	union	6		38	life	2
17	do	6		39	were	2
18	was	6		40	state	2
19	new	6		41	liberty	2
20	Public	6		42	time	2
21	constitution	6		43	spirit	2
22	such	5		44	justice	2

As can be seen in Table 4, most words are concentrated on the state and its citizens (e.g. *government, people, country, nation, America, union, public, citizens*). There are also several words connected to freedom such as *peace, freedom, free, democracy, or liberty* which comply with officially declared principles of USA. We can also see that adjectives among thematic words are positive (e.g. *great, new, free*); probably in order to ensure people that the new president will bring better future. There is just one word which expresses negative connotations – *war*.

Liu (2012) claims that the US presidential inaugural addresses consist of eight general parts. With the exception of salutation and other formalities such as announcing entering upon office or articulating sentiments on the occasion, Liu (2012) identifies following parts:

a) Making pledges – "The new president carries out this speech act to help the pubic with confidence in the new leader and his government." (Liu, 2012, p. 2410)
b) Arousing patriotism in citizens
c) Announcing political principles to guide the new administration – "The basic principles that all presidents swear to follow comprise American Constitution, union, freedom and democracy…" (Liu, 2012, p. 2410)
d) Resorting to religious power: "Every president will refer to God many times in his inaugural address as God is the common religious belief for nearly all Americans." (Liu, 2012, p. 2411)

As can be seen in Table 4, the thematic words comply with aforementioned themes. Only resorting to religious power do not fully correspond to *TW*, because *God* occurs only one time (Text 20, Lincoln).

5. Conclusion and Discussion

This study analyses the vocabulary richness (*MATTR*), text activity (*Q*), and secondary thematic concentration (*STC*) of US presidential inaugural addresses. We discovered that there is no obvious general tendency through the more than two centuries long history and the style of the speeches is rather influenced by personality of each president. We also found out that the aforementioned features are not relevant to the political affiliation. In these aspects our findings are different from those in Czech presidential speeches (cf. Čech 2014). On the other hand, we discovered that the addresses in wartime significantly differ in terms of secondary thematic concentration. Another difference was found in recession time where vocabulary richness is significantly higher.

To sum up, US presidential inaugural addresses seem to be mostly determined by individual style of each speaker but some important circumstances such as war or recession can affect the speech to some extent. Finally, it is necessary to say that this work is just a first attempt to analyse the US presidential addresses by the aforementioned indices. Therefore, more analyses must be done to support or reject our preliminary claims.

References

Busemann, A. (1925). *Die Sprache der Jugend als Ausdruck der Entwicklungsrhythmik*. Jena: Fischer.

Carranza, I. E. (2008). Strategic political communication: a leader's address to the nation. *Nueva Revista de Lenguas Extranjeras 10, 25–56.*

Čech, R. (2014). Language and ideology: Quantitative thematic analysis of New Year speeches given by Czechoslovak and Czech presidents (1949-2011). *Quality & Quantity 48, 899–910.*

Čech, R., Garabík, R., Altmann, G. (2015). Testing the thematic concentration of text. *Journal of Quantitative Linguistics 22, 215–232.*

Covington, M.A., McFall, J.D. (2010) Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics 17, 94–100.*

Kubát, M. (2014). Moving window type-token ratio and text length. In: Altmann, G., Čech, R., Mačutek, J., Uhlířová, L. (eds.), *Empirical Approaches to Text and Language Analysis: 105–113*. Lüdenscheid: RAM.

Kubát, M., Matlach, V., Čech, R. (2014). *QUITA - Quantitative Index text Analyzer*. Lüdensheid: RAM.

Lim, E.T. (2004). Five Trends in Presidential Rhetoric: An Analysis of Rhetoric from George Washington to Bill Clinton. *Presidential Studies Quarterly 32, 328–348.*

Liu, F. (2012). Genre Analysis of American Presidential Inaugural Speech. *Theory and Practice in Language Studies 2, 2407–2411.*

Matić, D. (2012) Ideological discourse structures in political speeches. *Komunikacija i kultura online 3, 54-78.*

Milička. J. (2013) *MaWaTaTaRaD* (software). Available at http://milicka.cz/en/mawatatarad/

Peters, G., Woolley, J.T. (2015). *The American Presidency Project.* http://www.presidency.ucsb.edu

Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B.D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., Vidya, M.N. (2009). *Word frequency studies*. Berlin/New York: Mouton de Gruyter.

Popescu, I.-I. (2007). Text Ranking by the Weight of Highly Frequent Words. In Grzybek P, Köhler, R. (Eds.) *Exact Methods in the Study of Language and Text*: 555–565. Berlin – New York: Mouton de Gruyter,

Savoy, J. (2010). Lexical Analysis of US Political Speeches. In: *Journal of Quantitative Linguistics 17, 123–141.*

Tuzzi, A., Popescu, I.-I., Altmann, G. (2010). *Quantitative Analysis of Italian Texts*. Lüdenscheid: RAM.

Appendix

#	Year	President	Types	Tokens	MATTR	Q	STC	Thematic words
1	1789	Washington	594	1431	0.73	0.55	0.004	HAVE, GOVERNMENT, MORE
2	1793	Washington	90	135	0.70	0.65	0.022	SHALL
3	1797	Adams	794	2322	0.69	0.48	0.008	PEOPLE, GOVERNMENT, NATIONS, MORE, COUNTRY
4	1801	Jefferson	679	1732	0.70	0.46	0.003	GOVERMENT, HAVE
5	1805	Jefferson	777	2168	0.72	0.60	0.011	HAVE, PUBLIC, WHO, CITIZENS, FELLOW, STATE
6	1809	Madison	521	1177	0.70	0.53	0.001	HAVE, BEEN
7	1813	Madison	518	1211	0.71	0.62	0.017	HAVE, WAR, BEEN
8	1817	Monroe	980	3379	0.71	0.51	0.013	HAVE, BEEN, GOVERNMENT, STATES, GREAT, HAS, OTJER, PEOPLE, UNITED
9	1821	Monroe	1195	4466	0.71	0.63	0.015	BEEN, HAVE, HAS, GREAT, STATES, WERE, OTHER, WAS,

								WAR, UNITED, MADE, CITIZENS, SUCH, HAD, GOVERNMENT
10	1825	Adams	961	2917	0.67	0.56	0.016	HAVE, BEEN, HAS, UNION, GOVERNMENT, RIGHTS, OTHER, COUNTRY
11	1829	Jackson	500	1128	0.70	0.44	0.001	PUBLIC, HAVE
12	1833	Jackson	474	1177	0.70	0.57	0.010	GOVERNMENT, PEOPLE, UNION, STATES, HAVE
13	1837	Buren	1252	3846	0.74	0.61	0.015	HAS, HAVE, BEEN, PEOPLE, WAS, INSTITUTIONS, GOVERNMENT, WERE
14	1841	Harrison	1799	8469	0.68	0.58	0.008	HAVE, POWER, HAS, PEOPLE, BEEN, CONSTITUTION, GOVERNMENT, WAS, CITIZENS, OTHER, STATES, EXECUTIVE, COUNTRY, GREAT, SPIRIT, MORE, CHARACTER, SUCH, LIBERTY, STATE
15	1845	Polk	1255	4814	0.69	0.56	0.013	GOVERNMENT,STATES,HAVE, UNION, HAS, BEEN, POWERS, PEOPLE,COUNTRY, CONSTI-TUTION, INTERRESTS
16	1849	Taylor	481	1091	0.71	0.44	0.010	SHALL,GOVERNMENT, COUNTRY
17	1853	Pierce	1114	3344	0.73	0.55	0.003	HAVE, HAS, POWER, BEEN, GOVERMENT
18	1857	Buchanan	889	2836	0.72	0.50	0.015	HAS, STATES, HAVE, SHALL, CONSTITUTION,BEEN, GOVERNMENT,PEOPLE, QUESTION, GREAT
19	1861	Lincoln	1005	3635	0.71	0.53	0.010	CONSTITUTION, HAVE, PEOPLE, UNION, STATES, GOVERNMENT, SHALL, SUCH, LAW, DO
20	1865	Lincoln	335	705	0.71	0.66	0.016	WAR, GOD
21	1869	Grant	466	1132	0.73	0.50	0.002	COUNTRY
22	1873	Grant	520	1339	0.71	0.60	0.003	HAVE, BEEN, COUNTRY, WAS
23	1877	Hayes	798	2489	0.69	0.50	0.007	COUNTRY, GOVERNMENT, HAVE, STATES, PUBLIC, POLITICAL, HAS, PEOPLE, GREAT
24	1881	Garfield	966	2987	0.71	0.59	0.010	GOVERNMENT, HAVE, PEOPLE, HAS, STATES, CONSTITUTION, BEEN, UNION, GREAT, LAW
25	1885	Cleveland	643	1691	0.67	0.47	0.012	PEOPLE, GOVERNMENT, PUBLIC, WHO, SHALL, CONSTITUTION
26	1889	Harrison	1299	4398	0.71	0.52	0.010	HAVE, PEOPLE, BEEN, WHO, STTES, HAS, SHALL, LAWS, PUBLIC, WAS
27	1893	Cleveland	794	2028	0.71	0.56	0.011	PEOPLE, GOVERNMENT, HAVE
28	1897	McKinley	1186	3972	0.72	0.52	0.008	HAS, PEOPLE, GOVERNMENT, CONGRESS, BEEN, GREAT, HAVE, COUNTRY, MORE, SUCH, WAS, PUBLIC

29	1901	McKinley	809	2215	0.72	0.59	0.005	HAS, GOVERNMENT, PEOPLE, HAVE
30	1905	Roosevelt	383	991	0.70	0.49	0.012	HAVE
31	1909	Taft	1372	5438	0.69	0.51	0.006	HAS, HAVE, GOVERNMENT, BUSINESS, SUCH, PROPER, LAW, CONGRESS, BEEN, OTHER, TARIFF, RACE
32	1913	Wilson	626	1712	0.69	0.61	0.023	HAVE, GREAT, BEEN, HAS, MEN, GOVERNMENT, HAD, JUSTICE, LIFE
33	1917	Wilson	523	1531	0.68	0.49	0.023	HAVE, OWN, MORE, BEEN, SHALL
34	1921	Harding	1117	3346	0.73	0.53	0.012	WORLD, JAVE, AMERICA, WAR, HAS, NEW, CIVILIZATION, GOVERNMENT
35	1925	Coolidge	1158	4056	0.70	0.52	0.010	HAVE, HAS, COUNTRY, GREAT, WHO, BEEN, PEOPLE, GOVERNMENT, MORE, WHAT, DO
36	1929	Hoover	1022	3766	0.67	0.45	0.009	GOVERNMENT, HAVE, MORE, PEOPLE, PROGRESS, PEACE, WORLD, JEUSTICE
37	1933	Roosevelt	709	1883	0.71	0.47	0.006	HAVE, NATIONAL
38	1937	Roosevelt	684	1823	0.72	0.57	0.020	HAVE, GOVERNMENT, PEOPLE, BEEN, NATION
39	1941	Roosevelt	490	1346	0.67	0.62	0.009	NATION, HAS, DEMOCRACY, HAVE, LIFE, SPIRIT
40	1945	Roosevelt	259	559	0.68	0.63	0.011	SHALL, PEACE, HAVE
41	1949	Truman	739	2283	0.70	0.50	0.022	WORLD, HAVE, NATIONS, PEACE, FREEDOM, PEOPLE, FREE, UNITED, MORE, PEOPLES, SECURITY, DEMOCRACY
42	1953	Eisenhower	845	2461	0.70	0.46	0.012	FREE, WORLD, PEACE, SHALL, HAVE, PEOPLE, STRENGTH, FREEDOM
43	1957	Eisenhower	585	1660	0.70	0.42	0.015	WORLD, NATIONS, FREEDOM, PEOPLE, PEACE, SEEK, OWN
44	1961	Kennedy	531	1365	0.70	0.48	0.012	LET, DO, WORLD, SIDES
45	1965	Johnson	524	1493	0.67	0.63	0.020	HAVE, NATION, CHANGE, MAN, UNION, WHO, PEOPLE
46	1969	Nixon	704	2131	0.69	0.57	0.011	HAVE, PEOPLE, WORLD, PEACE, WHAT, LET, WHO
47	1973	Nixon	505	1818	0.65	0.52	0.048	LET, AMERICA, PEACE, WORLD, HAVE, NEW, DO, HAS, RESPONSIBILITY, MORE, NATION
48	1977	Carter	491	1226	0.72	0.44	0.016	NATION, NEW, HAVE, HAD
49	1981	Reagan	841	2446	0.72	0.60	0.014	HAVE, GOVERNMENT, DO, WHO, HAS, BEEN, BELIEVE, AMERICANS, WORLD, PEOPLE
50	1985	Reagan	855	2575	0.73	0.59	0.018	HAVE, GOVERNMENT, PEOPLE, WORLD, FREEDOM, WHO, HAS
51	1989	Bush	743	2335	0.70	0.56	0.013	HAVE, NEW, WHAT, WHO, NATION, WORLD, GREAT
52	1993	Clinton	596	1611	0.71	0.54	0.034	WORLD, AMERICA, HAVE, PEOPLE, TODAY, WHO

53	1997	Clinton	717	2171	0.70	0.43	0.029	NEW, CENTURY, WORLD, AMERICA, NATION, HAVE, TIME, PEOPLE, LAND, GOVERNMENT, PROMISE
54	2001	Bush	583	1593	0.71	0.53	0.010	AMERICA, NATION, STORY, COUNTRY, CITIZENS, DO
55	2005	Bush	720	2078	0.70	0.57	0.024	FREEDOM, HAVE, AMERICA, LIBERTY, NATION, OWN
56	2009	Obama	886	2407	0.73	0.59	0.010	HAVE, HAS, WHO, NATION, NEW, AMERICA
57	2013	Obama	772	2120	0.72	0.58	0.003	PEOPLE, TIME

Glottometrics 34, 2016, 28 - 43

Intertextual Distance of Function Words as a Tool to Detect an Author's Gender: A Corpus-Based Study on Contemporary Italian Literature

Claudia Bortolato[1]

Abstract. The aim of this research is to explore the predictive capacity of function words in gender identification. In particular, from the different methods available the intertextual distance approach, as proposed by Labbè, was employed. The corpus analysed consists of 30 extracts from Italian novels, by the same number of authors, written between 1946 and 1967. Overall the corpus comprises over one million word tokens and more than 58,000 word types. The outcomes are in line with previous research that recognises function words as good indicators for the categorisation of writers according to their gender. However, the results of this study seem to be relatively weaker: the level of accuracy for women reached 63.33%, that of men 59.05%. In addition, the analysis of the category of function words as a whole turned out to have more predictive capacity than the two sub-categories considered, namely conjunctions and pronouns.

Keywords: *Author profiling, gender identification, intertextual distance, bag-of-words approach, function words, Italian literature.*

1. Literature background

Literature on language and gender has a long history of research. It is commonly acknowledged as dating back to Otto Jespersen, who discussed the topic in a chapter of his book 'Language: its nature, development and origin' (1922). Other pieces of research followed in the succeeding decades, among them the studies by Lakoff (1975), Maltz and Borker (1982), Tannen (1990, 1994) and Romaine (1994).

Their work, although stressing different aspects in terms of the causes behind gender language differences and the linguistic features displayed by men and women seems to share some traits: a more qualitative approach and a major attention to oral language over written texts. The analysis and discussion deals in particular with the sphere of lexicon: semantics, choice of topics and language in its empathic function.

More recently research has moved from this perspective to a different type of analysis. Alongside an interest in the studies of computer mediated language (e.g. Schler et al., 2006; Herring and Paolillo, 2006; Sarawgi et al., 2011; Mikros, 2013a - weblogs, Mikros and Perifanos, 2015 – tweets, Rangel and Rosso, 2013 - facebook), research has become more systematic, adopting a more quantitative approach and paying particular attention to methodological aspects. Thanks also to the development of the Natural Language Processing field, new algorithms have been made available to support the analysis. In terms of results, the

[1] University of Exeter, UK. Address correspondence to: claudia.bortolato@outlook.com

focus of the research field has shifted to include grammatical aspects. In particular, features related to morphology have eventually been seen to have good predictive power. This represents a huge revision in the perspective because elements that are semantically neutral are included – with promising results - in the detection of the writer's gender. The soundness of this approach is supported by the work of Mikros (2013b), among others, who found that males use more adverbs, coordinate conjunctions and contracted forms of prepositions whereas women more adjectives and personal pronouns; by Argamon et al. (2003), who observed that males use more determiners whereas females overall use more pronouns – although this last feature displays some exceptions; and by Rangel and Rosso (2013), who noticed a comparatively greater use of prepositions by men and comparatively more use of pronouns, determiners[2] and interjections by women. Sarawgi et al. (2011) have asserted that the strongest approach revolves around character level language models that detect morphological patterns, and that this approach is more robust than those examining lexico-syntactic patterns.

So far, results showing a difference in the use of language between men and women have been consistent in a wide range of corpora. Even avoiding for gender bias in topic and genre, evidence of a gender specific language style has emerged (Sarawgi et al., 2011). The genre of literature, which of its nature is a very personalised way of writing that embodies the individuality of its author, has turned out to have a relatively high level of gender attribution accuracy as well. The study by Koppel et al. (2002), comparing fiction and non-fiction corpora, found a result for fiction corpora, which are included in the literature category, ranging between 74.5 (when trained on both fiction and non-fiction) and 79.5% (when trained on fiction)[3], although non-fiction corpora performed better.

Discussion on the outcomes of different studies has highlighted some problematic methodological aspects and the need to adopt some caution. This applies while choosing our approach of analysis and research tools, and eventually discussing the outcomes, as well as when comparing results from different pieces of research. It has emerged, in fact, that instruments and methods of analysis are sensitive to the type of texts analysed (Tuzzi, 2010), and the specific questions addressed. Therefore they need to be carefully selected. For instance, approaches proving to be valid for gender detection may not be as successfully applicable to other studies, namely authorship attribution[4] (Mikros, 2013a); the gender of the author, in fact, is carried through certain syntactical and morphological patterns, whereas authorship attribution emerges from the under/over representation of certain high frequency words (Mikros, 2013a). In parallel the research by Koppel et al. (2002) showed that the training (algorithm) with mixed corpora, fiction and non-fiction, undermines the correct classification, and that non-fiction and fiction corpora follow different categorisation rules, as they differ in terms of the frequency of their distinguishing features. The study by Herring and Paolillo (2006) underlined the impact of the chosen approach as well, by pointing out how in order to obtain a sound interpretation of the results and to account for possible contrasting results with other studies, it is important to take into account the different methodological choices behind each work.

[2] What the authors call 'determinants'.

[3] As discussed by the authors the results depend also on the training set employed – only fiction on one hand or fiction and non-fiction on the other.

[4] For a summary of authorship attribution typologies see Savoy (2012).

2. Corpus

The corpus examined in this piece of research consists of 30 extracts from Italian literature written by an equal number of writers, evenly divided between men and women. From each novel about 35,000 words were taken, starting from the beginning of the book. Overall the corpus was composed of 1,057,753 word tokens and 58,518 word types.

A key aspect in building a corpus is the ability to make it as representative as possible of the phenomenon investigated. After selecting an independent variable – in this study it is the author's gender – it is then crucial to keep all the other variables at play constant; in this piece of research some variables related to the historical time to be considered, the social and educational background of writers and the type of texts. With regard to the historical time, all the books had to be written between 1946 and 1967. This thirty-year period is long enough to offer a sufficient number of texts to choose from, but it is still also circumscribed so that they can be relatively internally homogeneous. The span of time is in fact delimited by two major historical and cultural breakpoints: the Second World War and the protests of 1968. These events had an impact not only in both human and social terms, but also at a cultural and thus linguistic level.

With regard to the background of authors, the two main variables considered were their education and social background. At that time, the great majority of writers were likely to be of a higher social status and to have received a university education. There was still, in fact, a link between writers as professionals, a middle class background and a high level of education. Although this frame could be widely applicable and was particularly useful in keeping the other variables under control, it set extremely strict parameters and did not allow to be included people who for personal or historical contingencies had had to interrupt their university studies, or those who, though not from a middle class family, nevertheless were able to pursue their studies to a high level. Although, as said, the two variables of education and social background were largely shared by many writers, we decided to include some authors who, although not strictly meeting these criteria, were nonetheless active within the major cultural circles of the time. Furthermore, as the criterion of homogeneity pertains to the comparison between the two groups of female and male writers, we verified that authors who did not completely fit the educational/social criteria would have counterparts with similar social traits in the other group. One last criterion regarding the authors was 'age'. This operates on two levels: when it correlates with a time-frame, it helps to isolate people belonging to the same generation and therefore with a more similar education/culture. In general terms, this variable was applied to this study to level out any interference it could have had with regard to the use of language. The phase of maturity, which we saw as being between 25 and 55 years, was chosen and therefore all the texts considered in this study had to have been written when the author's age fell within that span of time.

With regard to the genre of texts, we decided on novels. The choice of a particular book from the range of possible options for a given author was casual. However, given the restrictions mentioned above, the range itself was limited: in a few cases there was even only one novel that could fit our criteria. Moreover the further requirement of 35.000 tokens from each novel excluded those below this minimum bar. Given the tight restrictions on the construction of the corpus, our choice was to obtain the books that met our criteria in paper format, and to scan the pages needed using optical character recognition (OCR) software. Since such scanning may include some typo errors, the texts were manually checked and corrected.

Other similar works controlled for the topic, regarding it as potentially impacting on the outcomes (e.g. Sarawgi et al., 2011; Mikros, 2013b). Therefore in those studies the topic

was kept consistent in both corpora, that produced by men and that by women. Although recognizing the grounds of this approach, topic was not strictly considered in this study. Firstly, taking into account this further variable would have made the selection criteria grid too strict and would not have allowed us to have a large enough corpus to use. As seen above, it also happened that in some cases there was just one book that met our criteria. The selection of female writers was made particularly difficult by their limited presence in the panorama of Italian literature of that time, when women were still somehow underrepresented in general publishing. Secondly, reducing the span of time considered allowed us to narrow down the topics covered in the novels and therefore to favor more homogeneity.

3. Research questions

The aims of this paper are to test and discuss different measurements for gender attribution in a corpus of 30 extracts from novels of post Second World War Italian literature. The approach of this analysis takes into account the perspective on authorship attribution "What to put in the bag" discussed by Tuzzi (2010), while applying it to gender identification. As Tuzzi (2010) pointed out, authorship attribution of literary texts is relatively robust compared to other genres, in particular when examining the inter-textual distance of function words. These latter have turned out to be powerful in isolating the personal style of an author, and therefore they represent a valid discriminator between writers. Thus, although aware that there are language choices that are personal traits shared with other people regardless of gender and that the approaches to gender identification may somehow differ from those for authorship attribution, in this present work we try to assess whether we can trace differences that are in some way persistent and detachable through statistical algorithms and eventually lead to a classification according to their authors' gender.

In this work we have in particular one research focus: considering previous studies, which have shown function words as powerful elements for - among other things - gender detection, we want to point the analysis to this category as a whole; namely prepositions, conjunctions, pronouns, determiners and adverbs[5], and also to look at some of its sub-categories separately, assessing which among them can better catch the gender of a text's author.

4. Automatic textual classification

One of the branches of computational stylometry is automatic author identification. This computer-based research field aims at identifying the author of a given text, by analysing a range of linguistic features. This approach is based on the idea that texts contain some sort of authorial fingerprint in the form of a selection of features that can discriminate between authors. Although the style itself cannot be directly measured, an indication of it can come indirectly from the occurrence and pattern of certain linguistic elements, which, on the other hand, can be quantified. To catch the stylistic elements peculiar to each author is not a straightforward task, and a result can be better obtained by looking at different indexes rather than opting for one cut-off element as a discriminator.

[5] The classification of the category of adverbs is open to debate and it is difficult to handle, as they may be considered partly as belonging to the category of function words and partly to that of content words.

On a broad level, when we start working on textual classification we need to address two steps: choosing both the linguistic features to analyze and the methods to establish the similarity/dissimilarity between texts.

4.1 Function words

As we have just mentioned, when approaching the topic of textual categorisation we firstly need to select the feature(s) to work on. Oakes (2014) underlines two traits that the elements taken into account need to display: they have to be frequent enough to support a statistical approach and become statistically evident, and they need to be objectively countable. Different studies have employed different features, ranging from specific grammar categories such as that of function words (e.g. Cortelazzo, Nadalutti and Tuzzi, 2012; Mikros 2013a), to character/word/POS n-grams (e.g. Koppel et al.2002; Mikros 2013a; Mikros and Perifanos, 2015) or vocabulary (Mikros 2013a and b). Some features may be, to a certain extent, tricky to handle and we have to be particularly cautious when examining the outcomes. As suggested by Oakes (2014) vocabulary richness, for example, is problematic as "it varies among authors and even within texts, so does not characterise the style of an individual author very well" (p. 6).

In this piece of research I have chosen to work with the category of 'function words'. They constitute a closed class group, meaning that their number is defined and cannot be increased. Their function is mainly grammatical, as they do not carry meaning but they serve as links between the different content elements within the sentence. As suggested by Stamatatos (2009), "function words are used in a largely unconscious manner by the authors and they are topic-independent" (p. 542). Therefore, given previous literature supporting their power in catching the style of authors – even related to gender classification – and the decision not to control the topic of the texts[6] and hence the need to reduce as much as possible the potential impact of the topic on our results, the choice of focusing on function words represented a more stable factor to help us detect an author's 'core' style across different books.

With regard specifically to this piece of research, I decided to work both with the category of function words as a whole, and to take into account a couple of its sub-categories, namely conjunctions and pronouns. The choice to restrict the analysis to this small number of sub-categories was driven by the properties of these two groups. Referring to the duality within the language between 'Grammatica delle regole' (grammar of rules) and 'Grammatica delle scelte' (grammar of choices), as discussed by Prandi (2006), I applied this concept in the study relating it to different categories of function words. Language is layered and it exhibits elements of formality, rigid repertoires of rules (grammar of rules), and elements of functionality, a range of discretionary options available to users (grammar of choices)[7]. In this respect, we can consider grammatical elements as closer to the sphere of rules than the lexical elements. However, it seems safe to say that even within grammatical words some sub-groups can be considered to be governed by rules more strictly, and, among them, conjunctions and pronouns seem to provide the user with more room for choice than other grammatical sub-categories. As a consequence they would allow the writer's stylistic individuality to emerge. Also, with reference to pronouns, this category has been previously linked to the social variable of 'class' (Bernstein, 1971), and so we may advance the hypothesis that the social

[6] I did however choose to consider only novels.

[7] This bi-dimensionality is partly blended. In particular Prandi (2006) underlines that there is an area (syntax) that is the result of both rules and choices.

rule that appeared to have an impact on the relation of 'language choices' and 'social class' may somehow operate also with regard to the social variable of 'gender'.

4.2 Intertextual distance, Purity index and Dendrograms

Intertextual distance is one of the techniques employed for text classification/clustering[8] (e.g. Labbé and Labbé, 2001; Labbé, 2007; Pauli and Tuzzi, 2009; Tuzzi, 2012). The notion of intertextual distance was developed from Muller's seminal work (Muller, 1968; Muller, 1977). The version, chosen for this analysis, is based on the one formulated by Labbé (Labbé, 2007; Labbé and Labbé, 2001) who proposes the use of intertextual distance as a tool to measure proximity/distance between large text corpora. The intertextual distance is based on the sum of the differences, in terms of frequency of the words, between two texts[9]. If we consider two texts, text A and text B, with N_A and N_B are their sizes, and where the dimension of text A (N_A) is smaller or equal to that of text B N_B , their intertextual distance is calculated as:

$$d(A,B) = \frac{\sum_{w \in V_{A \cup B}} \left| f_{i,A} - f^{*}_{i,B} \right|}{2N_A} \qquad (1)$$

where $V_{A \cup B}$ corresponds to the vocabulary of texts A and B and the frequency $f_{i,B}$ of each word in text B, the larger one, is scaled down according to the size of text A by means of the proportion $f^{*}_{i,B} = f_{i,B}N_A/N_B$ (Tuzzi, 2010; Cortelazzo, Nadalutti and Tuzzi, 2013). The value that the intertextual distance can assume is positive and it varies between 0 and 1; in the first case the two texts are identical, in the latter they share nothing (Savoy, 2015). In this work the 'size' of the vocabulary taken into account consists not of the whole vocabulary – as it was in the approach proposed by Labbè - but in a bag-of-words approach comprising a number of selected (function) words.

The last process of the analysis is performed by representing the outcomes in the form of a hierarchy-based dendrogram (tree), which can visually describe the relations between the objects (in this work between the thirty writers) and which is a valid tool to visualise a hierarchical clustering. For each set of words taken into consideration in this work, namely conjunctions and pronouns as well as for the whole category of function words, this procedure was replicated – with the aim of grouping texts - using a (cluster) algorithm with complete linkage method in order to group texts (Everitt, 1980). There are two types of hierarchical clustering: the agglomerative and the divisive. The one used here is the agglomerative, where each text represents a single cluster with just one observation in it. So if we have *t* texts we also have a *t* number of clusters with just one object in them. The distance between pairs of clusters is calculated as the maximum pairwise distance between each object of the two clusters. We then merge together the two most similar clusters, and we repeat the procedure until all the texts are combined into one single cluster. The dendrogram that we obtain visualises the distances between the clusters: the leaves represent the objects (in this analysis the thirty authors) and the branches represent the lines that join the objects into subgroups of

[8] For a review of measures that can be used to estimate the distance between two texts see Oakes (2014).

[9] As it is based on item frequency as a similarity/dissimilarity measure, the row ID index is dependent on the text size. Therefore it has to undergo adjustments when the texts are of different lengths (see Cortelazzo, Nadalutti and Tuzzi, 2013).

the closest texts. We can read the dendrograms horizontally, where the more branches the texts share the more similar they are. Ideally, as we are working with the variable of gender that has two modalities, the two clusters that we obtain before the last merging should each contain fifteen authors of the same gender. The closer we get to this result the more the clustering procedure proves itself to be accurate.

Another option to report the results is the discussion of the level of purity, as proposed by Tuzzi (2010). Once the intertextual distance has been calculated for all the pairs of texts, we can draw a square pair-wise matrix (*t***t*) of similarity measures between texts. This symmetrical matrix can be read as a ranking system: for each text, all the remaining *t* - 1 texts can be arranged in order, starting from the closest to the furthest, according to the distance value. If we compare the ranking that we obtain with the ideally expected classification we can assess the power of the procedure adopted to detect the similarity of our texts. If we consider a corpus of *p*+1 texts belonging to the same group (in this piece of work fifteen extracts from novels by the same number of female and male writers) we should ideally expect, for each text, all the remaining *p* texts in the first *p* positions. This indicates that the texts belonging to the same group are the closest. In this piece of work, once the distances from each novel have been calculated, the remaining fourteen texts of the authors belonging to the same gender group should occupy the first fourteen positions in the sorted list. If all texts belonging to the same group produce this result, then all *p(p*+1) comparisons would be assigned with ranks between 1 and *p*, thus obtaining a 100% level of purity. Conversely, the higher the number of objects, belonging to the same category that don't rank in the first *p* positions and progressively distance themselves from those positions, the more the procedure shows a lack of power in matching the correct ex ante classification. We can therefore read the purity index as the percentage of correctly positioned comparisons, that is, the number of comparisons of pairs of texts belonging to the same group with ranks between 1 and p.

5. Analysis and results

The first step that I undertook was to process the corpus using lexical statistical software. I chose the software Taltac[10], which is devised for the Italian language in particular and offers many different tools to perform linguistic analyses. Among these is the automatic annotation of the corpus, in which each word type is labelled according to its grammatical category. However there are forms[11] tagged as ambiguous, either as they belong to two or more grammatical categories (see homonyms/polysemes) or as they are not in the vocabulary (such as archaic forms or words with a different spelling). In order to improve the quality of the data, and therefore the soundness of the analysis itself, it was decided to partially reduce the number of unlabelled forms by carrying out their annotation manually. However the task was time-consuming, and thus the forms that had many grammatical occurrences, for which manual categorisation was unfeasible, were left unlabelled and did not enter the analysis. The software employed to perform the analysis is *R* (*R* Development Core Team, 2015), which is free software used for statistical analysis, available under the terms of the Free Software Foundation's GNU Genral Public License in source code form.

The results of our study are displayed in the following dendrograms (tree analyses). As I said, this format can visually represent the proximity/distance of different authors: every branch, in fact, acts as a separator, splitting or grouping together authors that share similar

[10] http://www.taltac.it/it/index.shtml

[11] This analysis is based on forms (e.g. Labbè and Labbè, 2001; Labbè, 2007), although other studies have been based on lemmas (e.g. Pauli and Tuzzi, 2009).

traits according to a given variable. Therefore, authors that are placed on branches that start from the same fork are more similar. In the graphs each author is reported with a letter preceding her/his name, either an 'F' or an 'M', to indicate their gender.

5.1 Intertextual distance

Conjunctions and pronouns

The first two categories that we are going to consider are conjunctions and pronouns. The function of conjunctions is to codify connections and to expand the core of a sentence, attaching new elements. The usage of conjunctions is linked to syntactical choices, and, therefore, it is correlated to the sentence length and also to a certain level of text complexity[12]. The choice of a syntactic style is not strictly controlled by rules and it is not related to the information conveyed. It is partly a personal choice on how to organise this information. Previous research on gender with regard to the use of conjunctions has found no statistically significant difference between female and male writers (Nicolau, 2015).

On the other hand, pronouns are words that refer to other elements within the discourse. Although their use is regulated by grammar rules, the choice of whether to use them or to explicate the referent has a certain margin of negotiability. This can leave room for some traits of an author's individuality to emerge. As we have hinted at above, the use of pronouns has been linked to social background, and in particular to social class (Bernstein, 1971). Bernstein proposed the idea that the structure of relationships within the family impacts on the communication choices of children. In particular he referred to the existence of two codes, one mostly typical of the middle class (elaborated code – where the reference to the context and its dynamics is made more explicit) and the other one of the working class (restricted code – context specific, in that the reference to the context is taken more for granted)[13]. This has implications also for the number of pronouns employed, which is greater in the restricted code. Given the historical period when the novels under consideration were written, during which the process of a greater inclusion of women within society was still at its outset, we can speculate on a possible difference in the use of pronouns among female and male writers. In particular we can parallel Bernstein's suggestions hypothesising that women, at the time enjoying less room to negotiate their roles out of the range of those already assigned to them by society, were more likely to take things for granted. This could have had an impact on the use of the language, and in particular they could differ from men in a greater use of pronouns. However this tendency could possibly be counterbalanced by the nature of this sample: the fact that I am working on literature, which by its very nature is quite personally-styled; and having all the writers from a culturally high level environment, could level off a possible difference, inhibiting the impact of the gender variable on the outcomes.
In the following two figures, Figure1 and Figure2, we show the dendrograms of conjunctions and pronouns respectively.

[12] Note that the readability index of a written text (e.g. the Gulpease or Flesh indexes) is linked also to the average length of a sentence.
[13] This difference in linguistic behaviour represents a general tendency among children of different social classes – subject also to the context in which the interaction takes place – and not a clear-cut element of separation between social classes.

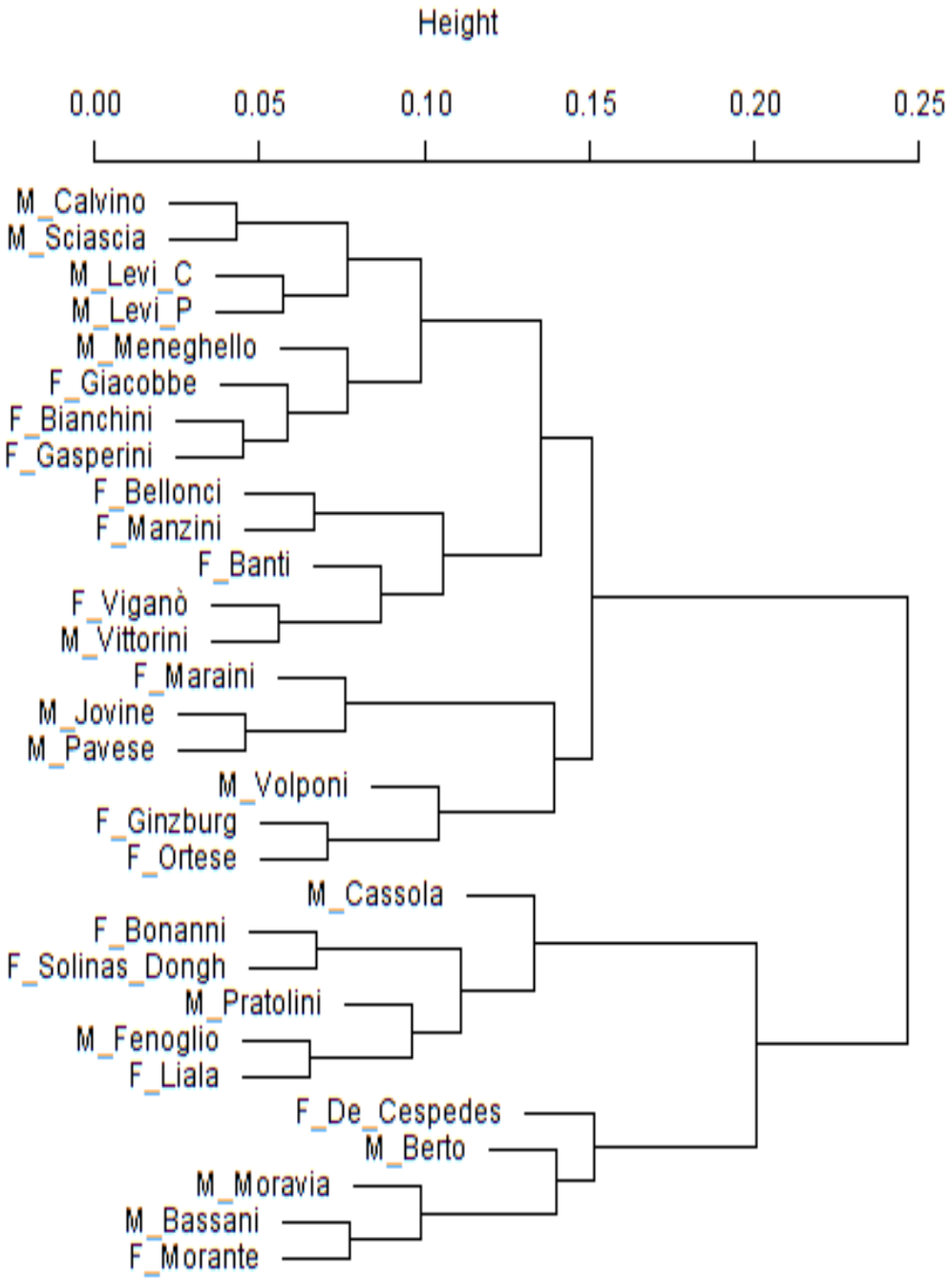


Figure 1. Intertextual distance and Conjunctions

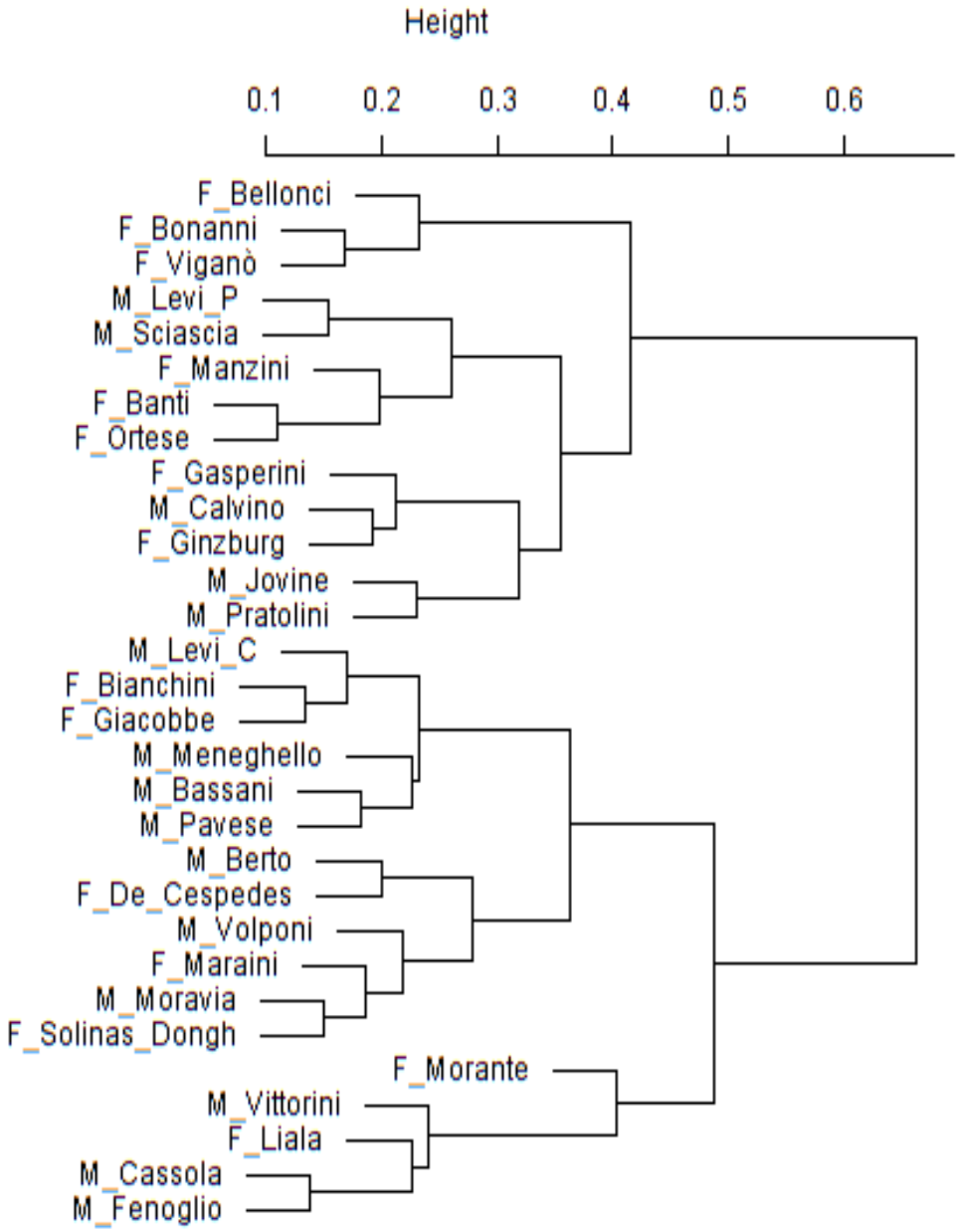


Figure 2. Intertextual distance and Pronouns

In both figures we observe no classification of the writers according to the variable of gender. However, as for the pronouns, the outcome seems to hint at some partition: as a result of the first ramification a small group of five authors – four of them female – are separated from the remainder. Yet, although the pronouns performed slightly better than conjunctions, their predictive capability seems nevertheless to be rather weak.

Function words

In Figure 3 the result for function words is displayed.

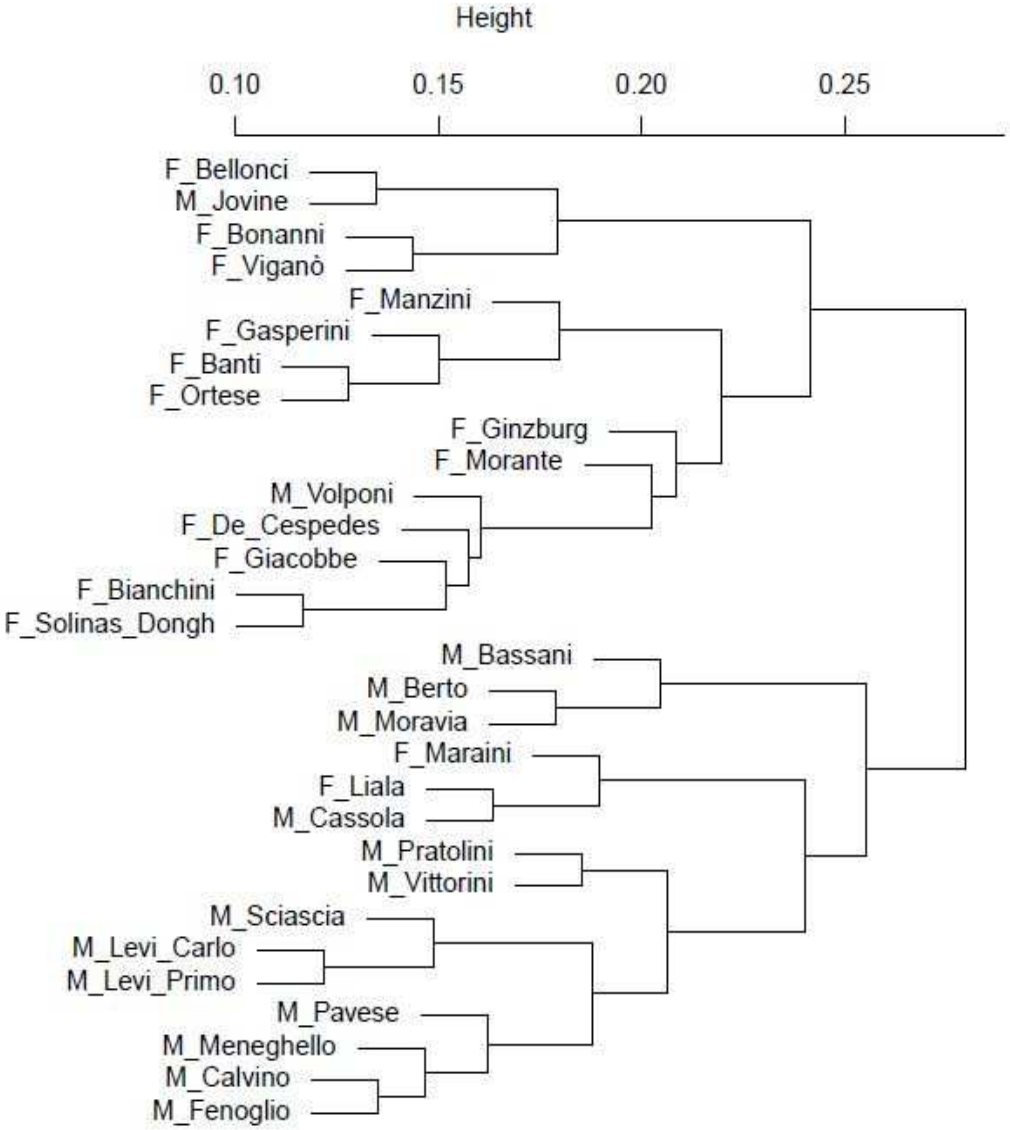


Figure 3. Intertextual distance and Function words

In contrast with the previous outcomes, the classification for the whole category according to the author's gender is confirmed overall. Only four authors, two from both groups, failed to meet the expected result; Maraini and Liala, were in fact classified in the group of male writers. Jovine and Volponi were, in contrast, placed with the female writers.

We lack an explanation for these four writers being categorised in the wrong groups, and research into both the books and the social traits of the authors does not seem to provide us with conclusive explanations. With regard to Liala and Maraini, in fact, both their books are about topics usually dealt with by female writers. Liala[14]'s book is a story of love, and

[14] Liala is best known as one of the main writers of romantic novels in the Italian literary panorama.

Maraini's one about a teenage girl. It seems somehow puzzling that the language of novels usually recognised as written by women for a female audience has been instead classified as having a male style. We do not have an explanation so far, although it is intriguing that even topics and writing style considered as more 'feminine' can adopt linguistic choices categoryised as male.

The occurrence of the two male writers categorised together with female writers is also unaccountable. Neither social nor autobiographical traits nor their topic choices seem to offer an explanation.

5.2. Purity index

Looking at the outcomes from a different perspective, I pass on to examine the level of purity. The results for this are lower compared to another similar piece of research on – among other genres - Italian literature (Tuzzi, 2010). The outcome for women (63.33%) is slightly higher than that for men (59.05%)[15]. With regard to female writers the range varies between 71% for five authors – Banti, De Cespedes, Gasperini, Maraini and Morante – and 50% for two others – Bianchini and Giacobbe. For male writers the level of purity ranges between the 71% of two authors – P. Levi and Vittorini – and the 43% of Moravia.

We can speculate that the relatively low outcomes may partially stem from the historical circumstances of the time in which these books were written. In particular, we may hypothesise that differences in writing style between female and male authors would be less pronounced at that time. Alongside a development in the topics that women in particular used to deal with in their writings (Bortolato, 2008), there could have been a counterpart in the development of their own feminine style. Up to the beginning of the last century, in fact, the very few female writers of some renown in the Italian cultural landscape used to work more on 'neutral' topics, addressing themes that were marginally related to the sphere of femininity or to themselves. A female literature that talks about women was a breakthrough of the twentieth century, in particular of its second part. This evolution could have possibly stemmed from the same social developments that have brought changes in women's writing style. Another related speculation regards the time frame that we set (see section 2). As we previously said, the Italian language did not undergo a regular development: it remained more or less stable for centuries and in the 20th century it entered into an accelerated process of transformation. The choice of a certain historical time, which was still more culturally and linguistically linked to the past – at least with respect to Italy – could be the basis of relatively lower results compared to others obtained working on more contemporary Italian corpora.

One last hypothesis pertains to the choice to employ literary texts. As we touched on above, we are dealing with texts characterised by a high level of individuality, regarding any social traits, and thus possibly fading in potential gender differences. However, despite this last hypothesis and the fact that the other variables have been kept under control as much as possible (see section 2), it shows up as of note that a difference, albeit slight, between female and male writers has still been detected. This follows other research, in particular Sarawgi et al. (2011), who suggested that indications of gender language styles are nevertheless detected even in a genre deemed to be less permeable to gender impact such as modern scientific papers.

[15] These outcomes are obtained by measuring the level of purity (as a percentage) for each writer and then calculating the means for men and women.

6. Conclusions

This piece of research aimed at testing function words as predictive elements for textual gender identification. The corpus we worked on comprises the novels of 30 Italian writers produced in the post Second World War period (1946 -1967). Given previous research on gender identification, and in general on author attribution, which highlighted the predictive capacity of the grammar category of function words, I decided to focus on this category, adopting a bag-of-words approach (see Tuzzi, 2010). In particular, I selected three categories to work on: conjunctions, pronouns and function words as a whole. With regard to the first two sub-categorisations, the outcomes revealed that these seem to have no clear predictive capabilities in identifying the gender of the author. On the other hand, the category of function words considered in its entirety improved the correct classification of the writers according to their gender. Overall these results seem to suggest that a more 'comprehensive' approach which includes more grammatical categories of function words is beneficial in improving the accuracy of the outcomes.

In line with the outcomes of research on similar topics, this study shows that function words are carriers of information containing some traces of the gender-genre, of the gender-sex and of an author's style. However we must proceed with caution in generalising the outcomes of this study: the corpus in fact consists of a portion (35.000 words) of a very limited number of texts (30) sharing very specific historical, social and literary characteristics. Besides, from a purely statistical perspective, working on intertextual distances based on a very limited number of forms, as is the case of function words, makes the analysis less stable. Sufficient to recall that the number of word types that entered the calculation reduces to a minimum of 13, 27 and 98 and to a maximum of 44, 43 and 152 with regard to conjunctions, pronouns and grammatical words respectively. Overall, intertextual distance seems a convincing instrument to measure the similarity between texts, and cluster analysis a valid method to represent the structure of a data matrix. However, at this stage, further measures to assess the stability of this approach would be necessary. For instance, we could perform randomisations, adding/removing one or more forms in the analysis, in order to verify the solidity of the adopted solution.

The combinations of a variety of procedures and of the variables taken into account, which widely differ among different pieces of research so far, make comparisons very thorny and not always reliable. As suggested by Savoy (2012), there are aspects regarding authorship attribution that are still critical and in need of development. In particular he indicates that, in evaluating corpora, "comparisons between reported performances and general trends regarding the relative merits of various feature selections, weighting schemes and classification approaches are difficult to assess with the required precision" (Savoy, 2012, p. 136). The lack of a clear definition of a shared research path, and a growing interest in the subject, make it desirable to devise a common framework to use in approaching this topic and thus the time seems ripe for proposing analysis schemes which can be shared among scholars. In this respect, the work by Juola (2015) represents a valid proposal which points out the pitfalls in approaching the topic, and suggests ways to define a shared work protocol within the community.

Other ways to deepen the analysis can take a more statistical stance, studying the interrelations between selected properties which may be different in texts written by female and male authors, i.e. by performing a reduced synergetic analysis and showing that the control cycles of the two groups have different parameters. Other ways can include taking into account further boundary conditions and applying other statistical methods. It may be

conjectured that it is possible to uncover other differences that can be expressed by a number of quantitative indicators.

Acknowledgments

I thank Prof. Ursini for having supervised the project from which this article developed, and Prof. Zambon for providing me with food for thought. I am especially indebted to Prof. Tuzzi for her competence and guidance, particularly for the statistical analysis.

Bibliography

Argamon, S., Koppel, M., Fine, J. and Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *Interdisciplinary Journal for the Study of Discourse 23(3), 321–346.*

Bernstein, B. (1971). *Class, codes and control* (volume 1). London: Routledge and Kegan Paul.

Bortolato, C. (2008). *Dimensioni di genere in scrittrici e scrittori del secondo dopoguerra. Un'analisi statistica.* Master thesis. University of Padua.

Cortelazzo, M., Nadalutti, P. and Tuzzi, A. (2012). Una versione iterativa della distanza intertestuale applicata a un corpus di opere della letteratura italiana contemporanea. In Diester, A., Longrée, D. and Purnelle, G. (eds), *Actes des 11es Journées internationales d'Analyse statistique des Données Textuelles: 295-307.* Bruxelles: Université de Liège - Facultés Universitaires Saint-Louis.

Cortelazzo, M., Nadalutti, P. and Tuzzi, A. (2013). Improving Labbè's intertextual distance: testing a revised version on a large corpus of Italian literature. *Journal of Quantitative Linguistics 20(2), 125-152.*

Everitt, B. (1980). *Cluster analysis*. New York: Halsted Press.

Herring, S. and Paolillo, J. (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics 10(4), 439 459.*

Jespersen, O. (1922). *Language: its nature, development and origin*. London: George Allen.

Juola, P. (2015). The Rowling Case: a proposed standard analytic protocol for authorship questions. In: *Digital Scholarship in the Humanities, (30),* Oxford: Oxford University Press.

Koppel, M., Argamon, S. and Shimoni, A.R. (2002). Automatically categorizing written texts by author gender. *Lit Linguist Computing 17(4), 401-412.*

Labbé, D. (2007). Experiments on authorship attribution by intertextual distance in English. *Journal of Quantitative Linguistics 14, 33-80.*

Labbé, C. and Labbé, D. (2001). Inter-textual distance and authorship attribution Corneille and Molière. *Journal of Quantitative Linguistics 8, 213-231.*

Lakoff, R. (1975). *Language and woman's place*. New York: Harper and Row Publisher.

Maltz, D. and Borker, R. (1982). A cultural approach to male-female communication. In: Gumperz, J. (ed.), *Language and Social Identity, 196- 216.* Cambridge: Cambridge University Press.

Mikros, G. K. (2013a). Authorship attribution and gender identification in Greek blogs. In: Obradović, I., Kelih, E. and Köhler, R. (eds.), *Selected papers of the VIIIth international conference on quantitative linguistics (QUALICO),* Belgrade: Academic Mind.

Mikros, G. K. (2013b). Systematic stylometric differences in men and women authors: a corpus-based study. In: Köhler, R. and Altmann, G. (eds.), *Issues in Quantitative Linguistics 3, 206-223*. Lüdenscheid: RAM – Verlag.

Mikros, G.K. and Perifanos, K. (2015). Gender identification in modern Greek tweets. In: Tuzzi, A., Benešová, M. and Mačutek, J. (eds.), *Recent Contributions to Quantitative Linguistics: 75-88.* Berlin, Boston: De Gruyter Mouton.

Muller, C. (1968). *Initiation à la statistique linguistique*. Paris: Larousse.

Muller, C. (1977). *Principes et méthodes de statistique lexicale*. Paris: Hachette.

Nicolau, M.F. (2015). Gender differences towards the use of conjunctions. Available at: http://www.academia.edu/15492327/Gender_Differences_towards_the_Use_of_Conjunctions

Oakes, M.P. (2014). *Literary Detective Work on the Computer*. Amsterdam/Philadelphia: John Benjamins Publishing.

Pauli, F. and Tuzzi, A. (2009). The end of year addresses of the presidents of the Italian Republic (1948-2006): discoursal similarities and differences. *Glottometrics 18, 40-51*.

Prandi, M. (2006). *Le regole e le scelte*. Torino: Utet Università.

Rangel, F. and Rosso, P. (2013). Use of language and author profiling: identification of gender and age. In: *Proceeding of the 10th workshop on natural language processing and cognitive science (Nlpcs)*. Available at: http://users.dsic.upv.es/~prosso/resources/RangelRosso_NLPCS13.pdf

Romaine, S. (1994). *Language in society: an introduction to* sociolinguistics. Oxford: Oxford University Press.

Sarawgi, R., Gajulapalli, K. and Choi, Y. (2011). Gender attribution: tracing stylometric evidence beyond topic and genre. In: *Proceedings of the fifteenth conference on computational natural language learning*, Available at: http://www.aclweb.org/anthology/W11-0310.pdf

Savoy, J. (2012). Authorship attribution: a comparative study of three text corpora and three languages. *Journal of Quantitative Linguistics 19(2), 132-161*.

Savoy, J. (2015). Authorship attribution using political speeches. In: Tuzzi, A., Benešová, M. and Mačutek, J. (eds.), *Recent Contributions to Quantitative Linguistics: 153-164*. Berlin, Boston: De Gruyter Mouton.

Schler, J., Koppel, M., Argamon, S. and Pennebaker, J. (2006). Effects of age and gender on blogging. In: *Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs: 199-205*. Menlo Parks (California).

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology 60(3), 538–556.*

Tannen, D. (1990). *You just don't understand*. New York: Ballantine Books.

Tannen, D. (1994). *Gender and discourse*. Oxford: Oxford University Press.

Tuzzi, A. (2010). What to put in the bag? Comparing and contrasting procedures for text clustering. *Italian Journal of Applied Statistics / Statistica Applicata 22(1), 77-94*.

Tuzzi, A. (2012). Reinhard Köhler's scientific production: Words, numbers and pictures. In: Naumann, S., Grzybek, P., Vulanovic, R., and Altmann G. (eds), *Synergetic Linguistics. Text and Language as Dynamic Systems: 223-242*. Wien: Praesens Verlag,

List of books by authors

Banti Anna	Artemisia
Bassani Giorgio	Il giardino dei Finzi-Contini
Bellonci Maria	I segreti dei Gonzaga
Berto Giuseppe	Il male oscuro
Bianchini Angela	Le nostre distanze
Bonanni Laudomia	L'imputata
Calvino Italo	Il barone rampante
Cassola Carlo	La ragazza di Bube
De Céspedes Alba	Dalla parte di lei
Fenoglio Beppe	Il partigiano Johnny
Gasperini Brunella	L'estate dei bisbigli
Giacobbe Maria	Diario di una maestrina
Ginzburg Natalia	Lessico famigliare
Jovine Francesco	Le terre del Sacramento
Levi Carlo	Cristo si è fermato a Eboli
Levi Primo	Se questo è un uomo
Liala	Ombre di fiori sul mio cammino
Manzini Gianna	Lettera all'editore
Maraini Dacia	L'età del malessere
Meneghello Luigi	I piccoli maestri
Morante Elsa	Menzogna e sortilegio
Moravia Alberto	La noia
Ortese Anna Maria	Il mare non bagna Napoli
Pavese Cesare	La luna e i falò
Pratolini Vasco	Cronaca familiare
Sciascia Leonardo	Le parrocchie di Regalpetra
Solinas Donghi Beatrice	L'estate della menzogna
Viganò Renata	L'Agnese va a morire
Vittorini Elio	Uomini e no
Volponi Paolo	Memoriale

Glottometrics 34, 2016, 44-55

Types of Hierarchies in Language

Gabriel Altmann[1]

Abstract. The article describes the necessity of modeling in terms of hierarchies which are parts of systems and shows some domains in which they can be found.

Keywords: hierarchy, levels, hypotheses, links, dependencies

1. Introduction

In standard linguistics, a problem was considered as solved if one found some rules, mostly grammatical ones. The great paradigms, structuralism and generative linguistics dominating in the 20th century, are the best examples. One forgot that rules are our own intuitional (!) constructs arising in the course of language evolution and are based on convention. Frequently, one obtained a classification based on crisp classes and considered the research as finished. Fuzzy and probabilistic thinking came secretly through the back door but it was easier and more secure to ignore it in order not to disturb the social peace. In the meantime, our big brothers, e.g. the physicists and biologists, entered dead and living matter and discovered that the way into the depth of their objects is infinite. What more, they began to converge and study the omnipresent information controlling the total reality. In linguistics, some researchers derived some laws and synergetic control cycles and stated that in language one has the same situation. There are properties linked by laws; that no property is isolated; that there is an overall self-regulation controlled by effort minimization. Today, we know that the number of properties of language entities is not finite but given by the state of the science, hence we can develop the control cycles ad infinitum. Besides, there is also another way on which at every stage one can construct new control cycles: that is the way into the depth of any entity (cf. Köhler, Altmann 1993)

In order to illustrate this way, consider for example the distribution of parts of speech in a language. Usually, one obtains 9-11 classes – as prescribed by the Latin grammar – but we know that the number of classes depends on the criteria we set up. If we study the frequency of these classes, we obtain a very regular rank-frequency distribution. But consider one of the word classes, say the adverbs alone. One finds again about 10 different adverb classes (depending on the official grammar). If one orders the adverbs found in a text (or in a corpus) in the prescribed classes, one can, again, search for the distribution of the adverbial classes. The simplest ordering is according to the rank-frequency of classes but one can devise a great number of ordering criteria e.g. from the semantic point of view, too. At this level one may ask whether the distribution of classes is law-like, what depends it on, etc.

Now, omitting all but one class (that is, concentrating to one sole class), one can study the logic of this unique class. If one considers e.g. adverbs of location, then location itself can be ordered. Some languages have special means for performing spatial orientation but even if the languages belong to the same family, translation shows that the semantics differs. How is the space oriented? Do the numbers obtained mirror this ordering? Is it possible to find a

[1] University of Bochum. Address correspondence to: RAM-Verlag@t-online.de

three-dimensional order? Does Man stay in the center? How is it with other adverbial classes? Can they be scaled or ordered in some way?

The next step is, again, to reduce the field to the given individual class and consider merely one of the adverbs. It occurs in different environments (= polytexty) and displays a polysemy which can be found partially also in translations of the pertinent sentences into various other languages. Again, find the rank-frequency distribution of the individual meanings. Is the rank-frequency distribution linked with some kind of possible scaling? What do we obtain if we transform the rank-frequency distribution in a frequency spectrum?

Now take the meaning represented by the most occurrences, i.e. the first in the ranking scale. Each occurrence may be realized in different contexts. Are all contexts identical or do some of them occur more or less frequently? Then set up the rank-frequency distribution of the polytexty of individual occurrences of the same meaning. Do it separately also for the other meanings. Find for all of them their rank-frequency distributions and show whether it is the same model or whether something changes when one goes to higher ranks? Can all of them be generated by the same mechanism with different parameters, or are they basically different? At this stage, even corpora can be used for obtaining data.

Up to now, we passed 5 stages, i.e. we made 5 steps into the depth. Is it possible that the same frequency regime rules at all stages? If so, what does change in the model in dependence of the depth? Necessarily, the parameters obtain different values, but perhaps some parameters must be added, some may be omitted. Can one set up, say, a differential equation in which the parameters can be interpreted as representatives of Köhlerian (2005) requirements?

The way is in no case finished. We have a rank-frequency of polytexties and take, say, the most frequent class. What kind of texts do we have? May we distinguish special classes of them? If so, then the given frequency can be represented as the sum of text type in which the adverb (having the given meaning) occurred.

But now we made a step to text type classification and reached a quite different domain. At each step in the hierarchy there are different steps possible according to our interest. Text sorts have properties which may be linked (or not) with those scrutinized by us. If we take – at whatever step – another frequency class, we may obtain a different result.

At last, there will be a net of links which will never be ready. This circumstance is caused not only by the extreme complexity of language but also by the fact that there are few linguists interested in this ladder into the precipice of our concept formation. The individual links must be derived and tested on many languages in order to obtain language laws. Unfortunately, there are too many languages and even a stepwise corroboration of a law-candidate is a very slow procedure.

Needless to say, rank-frequency is only one of the ways that can be gone. At every stage one can perform a scaling of the given property. Automatically one performs also measurements and may test a hypothesis derived from the general background theory. Usually, one obtains either a probability distribution or a simple function. From the epistemological point of view there is no difference because mathematical models do not express “truth” but make a “thing”, the object of investigation, more understandable for us, allow us operations within the model, allow us to link various phenomena. If we construct a set of interrelated hypotheses and test them positively, we can call them laws. We say then, that the subsumption of a phenomenon under this system of laws is our explanation.

Other kinds of explanation do not exist in linguistics. Rules (= conventions) do not explain anything but sometimes they can themselves be explained – or rather explicated – psychologically, culturally, etc. This is, of course, a very seldom event and is not part of linguistic theory. The description of circumstances is in itself no explanation.

The content of a linguistic theory is formed by some background mechanisms consisting of dependencies, or better, of links. They arise subconsciously, without conscious will

of language users a serve efficient communication. In the communication everybody cares for minimization of effort of any kind. Since the minimizations may be different, equilibrium must always exist, otherwise the communication would break down. The speaker and the hearer do not know it, they cannot set up linguistic laws or even learn to behave according to them; they merely abide by them, just as by Newton's gravitation.

Laws do not lie – as maintained once by a scientist – because they do not try to capture the "truth". Deriving hypotheses and corroborating them does not mean to have found the truth. The modeling by means of mathematics is merely our endeavor to make the reality understandable for our way of thinking and to make it formally operable. We know that other creatures classify the reality differently in order to find orientation and survive. Man developed mathematics which is merely a more exact expression of our views of reality – not the truth.

The way into the depth – that is, the way towards a theory – is manifold. We can climb downwards on a hierarchy tree but we can also extend our view at the same level/direction. Testing a phenomenon in ever more languages is one of the possibilities. Every new language/text may yield corroboration or rejection of our preliminary theory. In case of rejection one must check the data, their definition, find boundary conditions, if necessary, modify the hypothesis or modify the complete theoretical background. Especially the search for boundary conditions is a very complex procedure. Linguists frequently rely on the classification of languages (agglutinating, isolating, polysynthetic, etc.), even if we know that any such classification is merely a kind of categorical scaling. In systems theoretical presentation, a boundary condition is an addition of an influence in form of a constant, change of the existing link by replacing say, a constant by a linear function, etc. It is just in this point where mathematics can be of great help. It simplifies our way of thinking, allows us to express relationships symbolically, forces us to measure the properties and search for their links.

Language, being an overt dynamic system, must display at least one kind of hierarchy. But being the most abstract system which arose (= was created) "spontaneously", by self-organization, it is to be expected that it displays hierarchies of various kinds. As a matter of fact, we can find them everywhere as soon as we leave its surface and begin to enter some deeper levels. As a system, language does not differ either from the universe or from living organisms but its matter and structure are different. The common principles are: self-organization (evolution), self-regulation (striving for equilibrium), existence of subsystems (hierarchy), links between them, links between their elements, links between properties of elements, and the existence of environment. Each aspect is a subject of a separate discipline but the common core is the subject of synergetic linguistics. The number of aspects is not "given" by the reality but by the development of our science. From time to time new disciplines of linguistics arise or old ones get restructured (existence of "schools") and each step brings us nearer to considering the given aspect as system or at least to detecting some kind of hierarchy. The most famous examples are the dependence grammar and the generative linguistics whose sentence-trees are still in circulation but today one can evaluate them already quantitatively.

The simplest way to obtain a linguistic hierarchy is to use the formulation "*A* consists of *B, C,*…". Here *A* is the system, *B, C,*… are the subsystems or elements. Unfortunately, in qualitative linguistics one does not continue going into depth but stops somewhere at this level and shows that there are classes *B, C,*…, defined as … and differing from the other ones by … . Such procedures are necessary conditions for setting up a theory, but they are not sufficient. As a matter of fact, this is merely the beginning of science: concept formation. Needless to say, it is a very important beginning, because no science is possible without concepts which were coined by means of some criteria. Linguistic criteria do not exist in the material world, they are our concepts, too.

Detecting hierarchies can begin at any level and any entity. One can separate a hierarchy from other aspects but, as a matter of fact, all are linked in some way and a consequent continuation would always lead to a control cycle. However, partial investigations with clear beginning and end of the hierarchy can help to discover unknown connections.

The present article is merely a short survey showing the possibilities of setting up hierarchies in different domains of language and having a restricted scope. It is to be remarked that two different definitions of the same entity may lead to quite different types of hierarchies and links with other entities.

2. Material hierarchies

If a material linguistic entity (syllable, word, sentence,…), i.e. a system, is composed of smaller entities, i.e. subsystems, then the material (spatial or durational) substance of the components depends on the length of the construct. The law controlling this relationship is known as Menzerath's laws. Several relations displayed in Table 1 have already been tested and one obtained always corroborating results. The literature is very extensive (cf. Glottotheory 5(1), 2014, 121-123). Problems arise only if one skips a level and studies the change of the indirect composing entity, e.g. word length as a function of sentence length. Those cases are marked with "???" in Table 1. Linguistic laws are neither deterministic nor transitive, hence the dependence may disappear in such cases (cf. Grzybek 2010, 2011; Grzybek, Stadlober 2007; Grzybek, Stadlober, Kelih 2007; Grzybek, Kelih, Stadlober 2008).

Table 1
Some material construct-component relationships

Construct	**Measured in number of**	**Property of the component**
Hreb	Sentences	Sentence length
Sentence	Clause s	Clause length
Sentence	Words(???)	Word length (???)
Sentence	Syllables(???)	Duration of syllables (???)
Clause	Word	Word length
Word	Syllables	Duration of syllables
Word	Syllables	Syllable length
Word	Sounds	Duration of sounds
Word	Morphs	Morph length
Beat	Syllables	Duration of syllables
Beat	Sounds	Duration of sounds
Syllable	Sounds	Duration of sounds

Some researchers considered also the set of sememes (meaning set) in the word as a composing unit whose extent reduces with increasing word length. This is a quite natural consequence of meaning specification by adding e.g. affixes to the word or compounding it. Here two different aspects of language (form and meaning) are linked by means of Menzerath's law (c.f. Altmann, Beöthy, Best 1982; Rothe 1983; Fickermann, Markner-Jäger, Rothe 1984; Sambor 1984; Hřebíček 1990). It is to be expected that some variants of the law or similar derivations will hold also for other part-whole relations.

This hierarchy may be presented in form of a tree, but omitting a node, i.e. to skip a level, makes the given law invalid and perhaps a quite different one is valid, or some new boundary conditions must be taken into account as shown in the above mentioned references. Further, here always properties are concerned which are easily measureable.

As already said, these relations are not transitive. If the unit is, say, a word and its immediate component is syllable and there is a strong relation, then the results of this relation cannot be transferred to the "word – phoneme", even if "syllable – phoneme" is evident.

3. Dependence of syllable length on word length

For the sake of illustration we show one of the many hierarchic relations that have been studied very intensively, namely a material dependence. The best known hypothesis is: *The longer the word, the shorter are its syllables*. Word length is measured in terms of syllable numbers, syllable length in terms of sound/phoneme numbers. One considers always the mean syllable length resulting from all words of the same length. The hypothesis has been tested many times (Menzerath 1954; Rettweiler 1954; Gajic 1950; Altmann, Schwibbe 1989) and it gave the impetus for mathematization (Altmann 1980). At the beginning, one tries to capture the phenomenon by an appropriate formula. Unfortunately, there are a large number of languages and texts hence one is forced to apply different models without knowing the boundary conditions. But step by step one strives for unification in order to attain a concentrated image. In studying length relations a great number of models have been applied (cf. Best 1997), at last a unified model has been found, namely the Zipf-Alekseev function/distribution. (cf. Popescu, Best, Altmann 2014) defined as $y = Cx^{a+b \log x}$. In Table 2 one finds the fitting of the function to the mean length of syllables in Croatian words of length x, as they were measured by Gajic (1950: 97).

Table 2
Mean length of syllables in Croatian words

Word length (in syllables) x	Mean syllable length y	Computed values $\hat{y}$
1	3.46	3.46
2	2.67	2.66
3	2.32	2.36
4	2.20	2.21
5	2.21	2.13
6	2.06	2.09
7	2.00	2.03
a = -0.4403, b = 0.0856, C = 3.4593, R^2 = 0.9932		

It can be remarked that here also the simple Zipf-function had been sufficient ($R^2 = 0.96$).

3. Lexeme nets

In lexicology, the meaning of lexemes is usually defined by synonyms and/or hypernyms. If one considers the first hypernym of a word in the dictionary, one finds again a set of its synonyms and at least one hypernym. If one follows this way, one can state that a lexeme with very general meaning (e.g. "thing" or "system" or "organism" or "entity") stays at the top of the hierarchy and quite special ones at its bottom. But even for an individual lexeme there may be simultaneously various hypernymic paths to different general concepts. For example, the lexeme "letter" has several hypernyms. In this way, the simple tree-like hierarchy changes to a net of hypernymy relations. Such a net, called lexical net, has a number of properties that can be measured and there are already some derived relationships (cf. Hammerl 1987, 1988, 1989; Sambor 2005;

Sambor 1997; Sambor, Hammerl 1991; Schierholz 1989). Some problems are shown in Strauss, Fan, Altmann (2008). The law called Martin's law (cf. Altmann, Kind 1983) based on the first work in this domain (Martin 1974) has been modified several times (cf. Hammerl 1989; Schierholz 1989) and numerous aspects of this relationship have been found. Many images of lexical hierarchies of this kind can be found in Sambor (2005). Good sources are WordNets and monolingual dictionaries waiting for quantitative evaluation.

One can measure the height of the tree, its width, its complexity, density, etc. If several special words are used, a complex hierarchic net may arise whose properties are treated in every textbook of graph theory.

The scaling of elements in the net can be performed by the number of meaning units (semantemes). A very general word e.g. "thing" has surely a smaller number of semantemes than a very special word, e.g. "pencil". That is, the more specific is the meaning the higher is the degree of the word. The method can be applied also to texts – cum grano salis – but first each word of the text must be placed in a lexicological net. Unfortunately, this is a "pencil and paper" work, it cannot be performed mechanically.

The nets of the text lie on different levels hence they form a very complex hierarchy. Up to now there are no trials just to capture the lexematic interweaving of a text. The easiest way would, perhaps, be the use of "hrebs" but this problem is a task for the future. The presentation of nets of this sort could lead to the foundation of text typologies, it would enable us to study the development of children speech, of an author, of a different text-type classification, etc.

4. Morphology

Morphology is the study of word formation and composition. A word is composed of morphemes and the morphemes may possess not only material properties (length, duration, complexity, position, continuity, form,…) but also polysemy and various grammatical functions. They can have even homonyms, e.g. the English "-s" and "-`s" in spoken form are homonyms. The meanings or functions may be common to different morphemes, hence automatically a hierarchy arises. In synthetic languages, the net may be very complex because the individual morphs may be signals of grammatical categories, parts of speech, derivations, compositions and the morpheme itself can be a lexeme or not; morphemes may be discontinuous, etc. The morphological structure of a language presented as a graph may be a very complex net whose properties may be compared quantitatively with those of other languages. As a matter of fact, the classical language typology is merely a reduction of a hierarchy or hierarchies to one sole level.

In some languages also the parts of speech are morphologically marked. For example in Slavic languages the nouns are marked for case, gender and number. In each gender there are paradigms with different inflection for case and number. There are nouns which exist only in plural, etc.

The same procedure can be applied for the morphological characterization of a text. One may classify or scale the morphemes according to independence, polysemy, placing within the word, etc. and display the text as a hierarchy of morphemes. The lowest level consists of independent morphemes, the higher levels are occupied by the dependent ones according to their status. Though this is no step into the depth of language, it is a hierarchical morphological representation of a text. The study of many texts yields average values that can be compared interlinguistically, intertextologically, interdialectally, etc.

As can be seen below, morphology of a language is a hierarchical construction of means. Usually it is well analyzed for every language but texts have not been analyzed as yet.

5. Syntax

In different types of descriptive grammar, sentence is presented as a dependence tree. Or, the sentence can be presented as consisting of two main parts, say noun phrase and verb phrase which in turn consist of either general parts or concrete words. In this way one obtains a tree. In dependence grammar one begins with the verb and shows the dependence tree consisting of the rest of the sentence. Of course, the dependences may be presented also horizontally drawing an arrow from the directly dependent to its head. Many quantifications have already been performed (cf. Liu 2007; Köhler 2011) but not all consider the hierarchy in syntax.

Hierarchies may be constructed also by considering verb valency. One begins with a verb at the highest level and shows its valencies which are positioned at the second level.. A similar problem is that of predication or specification. A noun can have several types of predicates: verbs, adjectives, adverbial expressions, etc. which specify it. All of them can be classified into types. Taking one type, one can state its deeper levels, take again one of the elements, etc. Thus, syntax is only the first, surface level of sentence, consisting of rules.

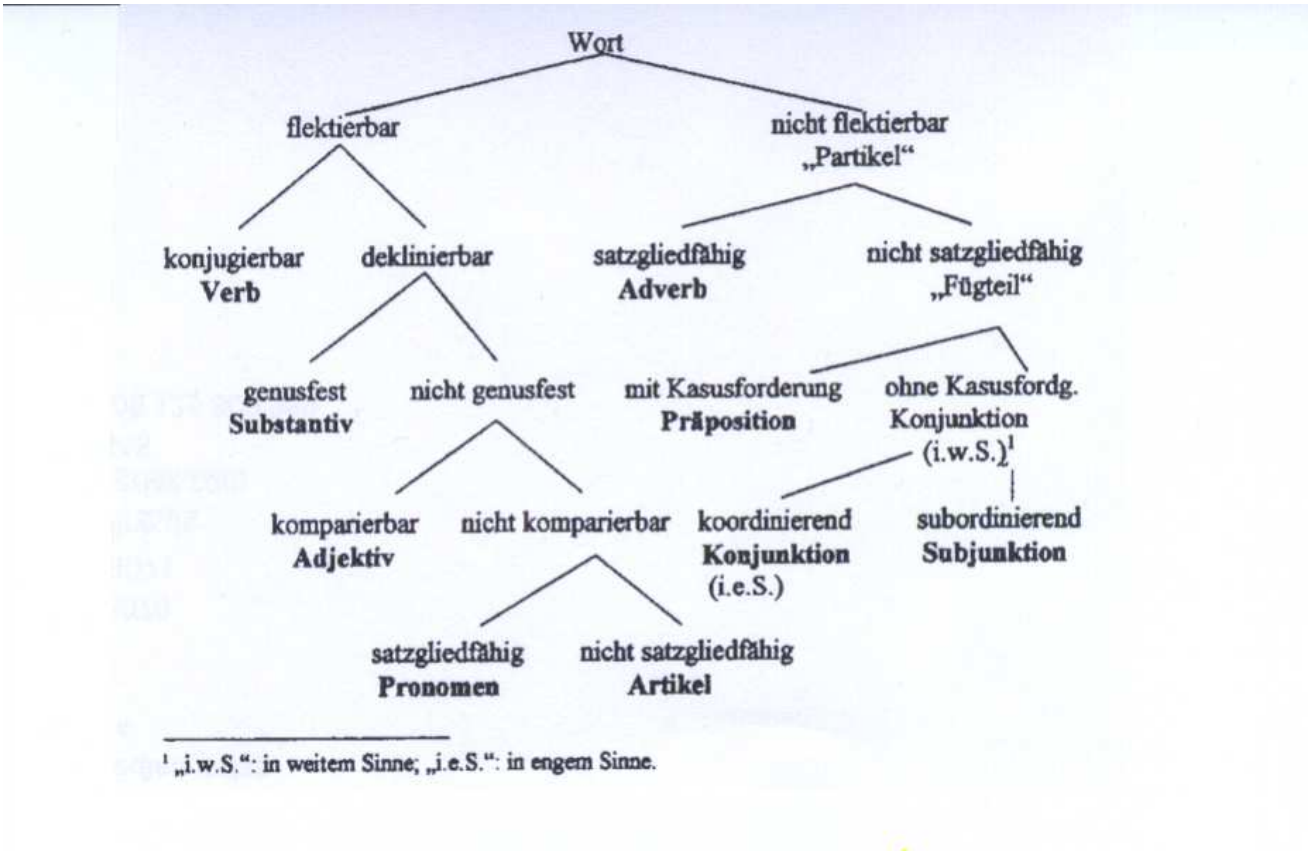
Texts can be analyzed hierarchically in the same way. The nets are that complex that the only way out of the labyrinth is the evaluation of some quantified properties of the graphs. Unfortunately, it has not been done frequently enough. The results of quantification of dependency or specification may be used for typologies, text type characterizations, study of the evolution of the respective aspects, etc.

6. Mixed hierarchies with word basis

In quantitative linguistics one frequently characterizes a text by the rank-frequency or the frequency spectrum of words (types, tokens) and tries to capture the distribution by a model. Sometimes one strives also for characterizing the stratification of the distribution and uses Popescu's approach by which the number of strata can be detected (cf. Popescu et al. 2010). If one discovers several strata, one is automatically at the top of a hierarchy. The strata may consist of synsemantics and autosemantics. However, even if one separates them and computes the stratification of one of the strata, one may obtain again several components. One can continue with classifying e.g. the autosemantics or all words in parts of speech according to the Latin tradition and compute the distribution of POS. Now, even if one obtains one component (or several ones), one can decompose the individual POS semantically. There are many analyses showing the semantic classes of nouns, verbs, adjectives; the classification of adverbs is taught also in grammar; synsemantics are "clearly" decomposed in prepositions, conjunctions, pronouns, articles, numerals, etc. each class having further subclasses. Within each class one can study the properties and classify the entities according to their value. This way has no end but it shows the depth of our argumentation.

Words can be classified in different ways. Since the number of word properties is not enumerable – it depends on the state of science and on our view – one can produce different types of classes. Some languages do it overtly by means of grammatical categories (e.g. gender; or noun classes in some South-East Asian and African languages marked with prefixes or numeratives; reflexivity marking, etc.) but mostly linguists do it because classes may have different properties, i.e. looking in the depth, classes may differ according to the view of the researcher. The best known classification in word classes is that of parts-of-speech inherited from ancient languages and surviving up to now in many grammars. Now taking a special class, we may find a sub-classification, and we may follow this way as long as necessary. Even if for grammatical description there is an end, for semantic description the end of the ladder is still deeper. Consider the hierarchy of word classes as defined for the German grammar (cf. Grundzüge 1980: 491; Best 2005: 41).

Now, at the end of each branch, there is a class of words which can be subdivided in various semantic or syntactic classes which in turn can be subdivided in phonemic classes. Somewhere the hierarchy ends, namely at the boundary of linguistic interests.



This hierarchy can be scaled very simply. Each word of the text can be placed at the given position and the position obtains as many scores as there are edges leading to it, counted from above. In this way, one can perform comparison of languages, e.g. translating a text, or study the development of language or of a writer. It is to be remarked that some of the classes have the same name (e.g. “satzgliedfähig” in the above figure) but they are scaled differently. The sentence can be transformed in numbers according to the above graph and the numbers can be processed statistically. Again, the network can further be refined if one defines classes within the given class.

7. Dialectology

No language is uniform in the entire territory where it is spoken. The official language has a number of dialects which have some common entities but other ones differ in various points. Even in the dialect one can recognize sociolects which associate a speaker with a special social class. And sociolects may be further composed of idiolects which bear personal features of the speaker and may be unique. The bearer of the idiolect has an active and a passive vocabulary and both can be further classified. Idiolects can be observed even in the same family because of different age and education of its members.

If we look in a dialect atlas, we see differently marked areas. The areas need not be continuous, the development may lead both to unification and to diversification depending on the ways of communication. The given language itself is not the highest level, because there are still subfamilies, families and maybe a common ancestor of all. Hence none of the two

directions (up and down) will ever bring a last form of the hierarchy, every site yields a different tree. Consider also the fact that every dialect is a mixture of vernacular words and borrowings which represent a hierarchy of the dictionary. Thus, from our point of view, a dialect is only a point of the hierarchy.

8. Frequencies

Even if frequencies (of phonemes, syllables, morphemes, words) seem to be numbers ascribed to elements of a class, there is a possibility to discover the number of strata on the given level using the Popescu method (Popescu, Altmann, Köhler 2010). It is not possible to identify the members of individual classes/strata directly but perhaps only after performing different partitionings of the inventory and finding the best simple functions. If one succeeds in attaining this aim, one can study the given sets separately, searching for the properties of individual elements. Thus the original set (say, of words) will be partitioned in two-three sets, and each of these sets will be in turn partitioned in further ones. Step by step one approaches very deep levels whose properties are linked with other properties. In this way one can discover quite hidden mechanisms of language which are not even mentioned in text-books.

Unfortunately, up to now there are not even trials in this direction, because one does not have criteria for constructing classes of components.

9. Hrebs and chains

If we study a written text, we find conventional punctuation which is not present in spoken language. But we may ask whether the sentence – there are about 200 definitions – is the highest level. Of course, one finds paragraphs and chapters, verses and strophes, etc. but we know that all these phenomena were created consciously by the author for some reason. Of course, one can consider them as units and insert them in the textual hierarchy. However, there are also entities that have been discovered recently, namely Hrebíček's (1997, 2000) *hrebs* and Belza's (1971) *chains* showing the semantic structure of the text and forming a supra-sentence level. Hrebs can be considered in various forms: as a set consisting of morphemes, words, phrases. A hreb comprises all sentences containing an element of the set. Each sentence of a text may belong to several hrebs at the same time. In this way the text can be presented as a graph which has a number of properties. One automatically asks whether there are still higher levels between the hreb-structure and text itself and we are persuaded that later on some will be found, it depends only on the view of the researchers.

Belza chains are, so to say, special restricted cases of hrebs lying in immediate neighborhood. A chain consists of subsequent sentences in which the same thematic word is repeated. A text contains chains of different length. A word may occur, say, in the first three sentences (length 3), then somewhere in the text in five subsequent sentences, etc. The distribution of chain lengths is a view of the semantic concentration of text. The chain need not contain the same word, it can be a synonym or a reference. If one numbers the sentences and joins those belonging to the chain, one obtains a graph which need not be completely connected but the degree of connectivity can, again be used as an indicator of thematic concentration.

Evidently, one may device different views of sentence cohesion or conceptual inertia (cf. Chen, Altmann 2015) which may serve as one level in a hierarchy.

10. Motifs

The usual Köhlerian motifs are non-decreasing sequences of numbers representing some property of the given entities in a text. They are secondary entities, i.e. they stay hierarchically above the given property. If we observe the length of words, then motif is a unit (sequence) made up of non-decreasing word lengths. The hierarchy continues in the same way as that of the original entities: there may be phrase motifs, clause motifs, sentence motifs, etc.

Motifs have a certain length, say, the motif [1,1,2,5] has length 4. If one counts the individual lengths, one obtains the distribution of motif lengths. Again, the distribution has its properties, etc.

There are also qualitative motifs as defined by Köhler, Naumann (2008). Hence, any sequence of some defined entities can be transformed in more abstract entities. The hierarchy does not, of course, end at this level because a sequence of motifs can further be reconstructed in, say, "super-motifs". As has been shown several times, motifs display regularities. Hence one can go into the depth on this way using entities which were not yet defined in linguistics.

11. Forming of hierarchies

Wherever one begins, every level of language takes a place in a hierarchy. The discovery of levels has the same importance as that in physics, but in linguistics one does not have a technical apparatus performing prescribed operations. The only possibility is relying on our concept formation and on searching for entities belonging (at least probabilistically) to the given subset. Many levels are already known but there are unknown ones whose effect cannot be observed in communication. Though human language is a human invention, its theory cannot be achieved merely by describing and classifying the evident phenomena or grammatical rules positioned at the surface of language, one must investigate thousands of ways into the depth, find links and dependencies, show the functioning of the open dynamic system called language. The search for hierarchies is one of the ways associated with much work, use of many data and a lot of mathematics.

The study of hierarchies is a presupposition for theory formation. If we consider language a system, then the levels cannot be ignored. But the elements of each level have a great number of properties which are linked with one another. Links of this kind are members of a self-regulation cycle. If a property changes, then other ones, connected with it, must change, too, otherwise the communication is endangered. The derived, well corroborated and substantiated links are the entities called laws.

If one has established hierarchies, one can use them as scales for the given property. The definition of a scale allows us to perform quantifications. Having hierarchies, the "higher" entities have a higher value on the given scale; having a set of entities on the same level, one can order them according to a different property. As a matter of fact, there is always a quantification possibility. The next steps are then the statement of hypotheses, sampling of data, testing the hypotheses and constructing step by step a control cycle which is a necessary condition for a theory.

References

Altmann, G., Beöthy, E., Best, K.-H. (1982). Die Bedeutungskomplexität der Wörter und das Menzerathsche Gesetz. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 35, 537-543.*

Belza, M.I. (1971). K voprosu o nekotorych osobennostjach semantičeskoj struktury scjaznych tekstov. In: *Semantičeskie problemy avtomatizacii informacionnogo potoka: 58-73.* Kiev.

Best, K.H. (ed.) (1997). *The distribution of word and sentence length*. Trier: WVT.

Chen, R., **Altmann, G.** (2015). Conceptual inertia in texts. *Glottometrics 30, 73-88.*

Fickermann, I., Markner-Jäger, B., Rothe, U. (1984). Wortlänge und Bedeutungskomplexität. In: Boy, J., Köhler, R. (eds.), *Glottometrika 6, 115-126.*Bochum: Brclmeyer.

Gajic, D.M. (1950). *Zur Struktur des Serbokroatischen Wortschatzes. Die Typologie der Serbokroatischen mehrsilbigen Wörter*. Bonn, Diss.

Grundzüge der deutschen Grammatik (1981). Berlin: Akademie.

Grzybek, P. (2011). Der Satz und seine Beziehungen. I: Satzlänge und Wortlänge im Russischen (Am Beispiel von L.N. Tolstojs «Анна Каренина») In: *Anzeiger für Slavische Philologie, 39, 39-74.*

Grzybek, P. (2010). Text difficulty and the Arens-Altmann law. In: P. Grzybek, E. Kelih, J. Mačutek (eds.), *Text and Language. Structures · Functions · Interrelations. Quantitative Perspectives: 55-70*. Wien: Praesens.

Grzybek, P., Kelih, E., Stadlober, E. (2008). The relation between word length and sentence length. An intra-systemic perspective in the core data structure. *Glottometrics 16, 111-121.*

Grzybek, P., Stadlober, E. (2007). Do we have problems with Arens' law? A new look at the sentence-word relation In: Grzybek, P., Köhler, R. (eds.), *Exact Methods in the Study of Language and Text: 205-217.* Berlin, New York: Mouton de Gruyter.

Grzybek, P., Stadlober, E., Kelih, E. (2007). The relationship of word length and sentence length. The inter-textual perspective. In: Decker, R., Lenz, H.-J. (eds.), *Advances in Data Analysis. Proceedings of the 30th Annual Conference of the Gesellschaft für Klassifikation e.V., Freie Universität Berlin, March 8-10, 2006: 611-618.* Berlin, Heidelberg: Springer.

Hammerl, R. (1987). Untersuchungen zur mathematischen Beschreibung des Martingesetzes der Abstraktionsebenen. In: Fickermann, I. (ed.), *Glottometrika 8, 113-129***.** Bochum: Brockmeyer.

Hammerl, R. (1989). Neue Perspektiven der sprachlichen Synergetik: Begriffsstrukturen – kognitive Netze. In: Hammerl, R. (ed.), *Glottometrika 10, 129-140.* Bochum: Brockmeyer.

Hammerl, R. (1989). Untersuchung struktureller Eigenschaften von Begriffsnetzen. Hammerl, R. (ed.), *Glottometrika 10, 141-154.*

Hřebíček, L. (1990). Menzerath-Altmann's law on the semantic level. *Glottometrika 11, 47-56.*

Hrebíček, L. (1997). *Lectures on Text Theory*. Prague: Oriental Institute.

Hrebíček, L. (2000). *Variation in Sequences***.** Prague: Oriental Institute.

Köhler, R. (2005). Synergetic linguistics**.** In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 760-774*. Berlin/New York: Mouton-de Gruyter.

Köhler, R. (2011). *Quantitative Syntax Analysis.* Berlin-Boston: de Gruyter.

Köhler, R., Altmann, G. (1993) *Begriffsdynamik und Lexikonstruktur*. In: Beckmann, F.; Heyer, G. (ed.), *Theorie und Praxis des Lexikons*: *173-190*. Berlin: de Gruyter.

Köhler, R., Naumann, S. (2008).Text Analyses Using L-, F-, and T-Sequences. In: Preisach, B., Schmidt-Thieme D. (eds.), *Proceedings of the Jahrestagung der Deutschen Gesellschaft für Klassifikation 2007 in Freiburg*: *637-646*. Berlin-Heidelberg: Springer.

Liu, H. (2007). Probability distribution of dependency distance. *Glottometrics 15, 1-12.*

Martin, R. (1974). Syntaxe de la definition lexicographique; etude quantitative des definissants dans le Dictionnaire fundamental de la langue francais, In: David, J., Martin, R. (eds.), *Statistique et linguistique*. Paris: Klincksieck.

Popescu, I.-I., Altmann, G., Köhler, R. (2010). Zipf's law – another view. *Quality and Quantity 44(4), 713-731*.

Popescu, I.-I., Best, K.-H., Altmann, G. (2014) *Unified Modeling of Length in Language*. Lüdenscheid: RAM-Verlag.

Rothe, U. (1983). Wortlänge und Bedeutungsmenge. Eine Untersuchung zum Menzerathschen Gesetz an drei romanischen Sprachen. In: Köhler, R., Boy, J. (eds.), *Glottometrika 5, 101-112.* Bochum: Brockmeyer.

Sambor, J. (1984). Menzerath's law and the polysemy of words. In: Boy, J. Köhler, R. (eds.), *Glottometrika 6, 94-114.* Bochum: Brockmeyer.

Sambor, J. (ed.) (1997). *Z zagadnień kwantytatywnej semantyki kognitywnej*. Warszawa: Polskie Towarzystwo Semiotyczne.

Sambor, J. (2005). Lexical networks. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 447-458.* Berlin/New York: Mouton-de Gruyter.

Sambor, J., **Hammerl, R.** (1991). *Definitionsfolgen und Lexemnetze, Band I.* Lüdenscheid: RAM-Verlag.

Schierholz, S. (1989). Kritische Aspekte zum Martinschen Gesetz. In: Hammerl, R. (ed.), *Glottometrika 10, 108-128.* Bochum: Brockmeyer.

Skorochod'ko, E.F. (1981). *Semantische Relationen in der Lexik und in Texten.* Bochum: Brockmeyer.

Strauss, U., **Fan, F., Altmann, G.** (2008). *Problems in Quantitative Linguistics Vol 1.* Lüdenscheid: RAM-Verlag.

https://en.wikipedia.org/wiki/Hierarchy

Glottometrics 34, 2016, 56-78

Interview with Jean Petitot

Jacqueline Léon, Sylvain Loiseau[1]

Abstract. From the 1960s onwards, the mathematician René Thom (1923-2002) carried out important contributions to the mathematical modelling of morphogenesis (analysis of forms). The proposed concepts (singularity, structural stability, catastrophe, bifurcation) were re-used in several social sciences, particularly in linguistics. They allowed a Gestalt-like approach, in opposition to the then dominant logico-combinatorial ones, and met some cognitivist trends and connectionist models. Jean Petitot was the first to show interest in the application of Thom's work to linguistics and he developed many studies accordingly.

This article is based on an interview between J. Petitot, J. Léon and S. Loiseau held on the 27 of September, 2014. While preserving the oral style of free conversation, it also includes references, developments and mathematical statements added by the authors.

I. Jean Petitot's biography

After preparatory classes at Louis-le-Grand high school, I entered the École Polytechnique in 1965 where I graduated in 1968. Impassioned by research, I joined the new Centre of mathematics my professor Laurent Schwartz had just created, and I learned algebraic geometry (with Grothendieck's disciple Jean Giraud) and differential geometry. As I investigated singularity theory[2], I met René Thom, one of the leading specialists of the field, who had restructured it entirely since the middle of the 1950s[3].

Besides, I was much interested in structuralism, in particular Claude Lévi-Strauss, whose lectures at the Collège de France I already attended when I was very young, around 18-19 years old. It is through Lévi-Strauss that I discovered Roman Jakobson. At that time, I did not see any link between structuralism and mathematics.

At the end of the 1960s, René Thom started to circulate the manuscript of *Stabilité structurelle et morphogenèse* (published in 1972). This book, which was focused in biology, also explained why the scope of the morphogenetic approach[4] went far beyond biology and could apply to structuralism in general. Thom had much discussed with Conrad Hal Waddington (1905-1975), an eminent specialist of embryogenesis, and he had benefited from Jakobson's strong support. His book, proposing a mathematical structural approach of biology, became very controversial. Its media audience owed much to Christopher Zeeman

[1] Université de Paris. Address correspondence to: sloiseau@u-paris10.fr

[2] cf. *infra*.

[3] René Thom (1923-2002) was a mathematician and a former student of the Ecole Normale Supérieure. He had lectured in Grenoble, Strasbourg and, near Paris, at the Institut des Hautes Etudes Scientifiques, until 1990. He received the Fields medal in 1958 for his work on differential topology. He developed mathematical models of morphogenesis popularized under the name of "Catastrophe theory" (cf. *Stabilité structurelle et morphogenèse*, 1972, InterÉditions, Paris).

[4] Morphogenesis studies the formation processes of complex forms, in particular those of life.

(University of Warwick), who coined the term "catastrophe theory" and turned it into a very general methodology whereas Thom's objectives were more focused[5].

In 1969 I discussed with René Thom his applications of singularity theory to structuralism. According to what he told me, I was the first young mathematician of my generation to do it. These discussions filled me with enthusiasm. After having hesitated between pure mathematics (I then had a position at CNRS, Centre National de la Recherche Scientifique) and modelling, I accepted in 1971 a position at the Center of Analysis and Social Mathematics (CAMS) of the 6th section of the EPHE (École Pratique des Hautes Études), which would become later the EHESS - Ecole des Hautes Etudes en Sciences Sociales). I was recruited at the EPHE thanks to the support of Lévi-Strauss, Fernand Braudel, the director of the 6th section, who believed in the role of mathematics in the social sciences, and Charles Morazé, my former professor at the Ecole Polytechnique.

Having joined the EPHE, I naturally got in touch with some structuralists, including A.J. Greimas. Greimas made an announcement in his seminar and some young colleagues got interested. I thus met Jean-François Bordron, Frédéric Nef, Paolo Fabbri, Jean-Claude Coquet, Per Aage Brandt, François Rastier, Claude Chabrol, and later Jacques Fontanille, Ivan Darrault, Jean-Jacques Vincensini and several other semioticians. Greimas did not have a very strong institutional position and his disciples had rather difficult careers, but he compensated for this fragility by his exceptional dimension.

Thanks to Paolo Fabbri who put me in touch with Umberto Eco, I spent one year in Bologna where I wrote a part of my Habiltation thesis ("thèse d'état"). I spread Thom's work on structuralism in the international semiotics community, in Bologna of course, then within Per Aage Brandt's group in Aarhus in Denmark and in Toronto. I very early discussed with Jean-Pierre Desclés who worked at the time with Antoine Culioli. Thus, I became involved in a social sciences community where I could use my double competence in mathematics and semiotics.

Later on, I remained primarily at the CAMS. I defended my "thèse d'état" in 1982[6], and, until 1985, I remained focused on applications of Thom's morphodynamical models (*i*) to phonetics (*Les catastrophes de la parole. De Roman Jakobson à René Thom.* Maloine, Paris, 1985), (*ii*) to elementary structures in semiotics, (*iii*) and to theories of actantial syntax, in particular case grammars (*Morphogenèse du sens. Pour un schématisme de la structure.* PUF, Paris, 1985).

Afterwards, I became more and more interested in cognitive neurosciences. In 1986, I joined a team of cognitive sciences which had just been created by Daniel Andler in a lab of the Ecole Polytechnique, the CREA (Centre de Recherche en Epistémologie Appliquée) founded in 1982 and directed by Jean-Pierre Dupuy. A little later, Michel Imbert, a specialist in neurosciences, created the first DEA (Diplôme d'Etudes Avancées - a post-graduate diploma) of cognitive sciences, and I became actively involved there in this new context, which led me establishing footbridges with American, in particular Californian, cognitivism.

[5] Christopher Zeeman (born in 1925) founded the Department of Mathematics and the Research Centre in Mathematics of the University of Warwick in 1964. In 1969-1970, during a sabbatical year in Paris, he discovered René Thom's Catastroph theory. Next, he largely contributed to the notoriety of this theory by providing it with many applications in various fields, in particular in social and behavioural sciences (cf. Isnard C. A. & Zeeman E. C. (1976). Some models from catastrophe theory in the social sciences". In: Collins L. (ed.) *Use of Models in the Social Sciences*, Tavistock, London, pp. 44-100).

[6] *Pour un Schématisme de la Structure: de quelques implications sémiotiques de la théorie des catastrophes*, thèse d'état defended in 1982 at the École des Hautes Études en Sciences Sociales, Paris.

II. René Thom's contributions

Q.: You said you were initially interested in Thom's work in mathematics, in particular in his work on the theory of singularities you were working on. Then, you were interested in his work on structuralism in linguistics. Can you tell us about Thom's contributions in these fields? To start with, what is his theory of singularities?

Singularity theory aims to study, analyze and classify geometrical structures of a specific type, which are called "singular" because they are not "regular". One considers a class of objects for which (*i*) the opposition between local and global properties is meaningful, and (*ii*) there are standard "simple" objects whose structure is "trivial". Then, one calls "regular" the objects that are everywhere locally simple (or "locally trivial"), even if they can be globally very complex and not trivial at all. There exist singularities when, locally, the considered object is not regular.

For example, let us consider surfaces and define a regular point as a point where the surface admits a tangent plane. Let us take a cone: apart from the vertex there is a tangent plane at every point and the cone is thus locally regular. But the vertex does not have a tangent plane and is thus a singular point. And as it is the only singular point in a small neighbourhood, it is said to be an isolated singularity.

One is immediately confronted with a theoretical problem: how to classify the singularities? One can observe for example that there exist points that are more or less singular. Consider for example a roof: apart from the ridge, points are regular. The points of the ridge are singular but are not isolated singularities since the whole ridge consists of singular points. The cone apex is more singular, it has a larger "degree" of singularity than the ridge of a roof.

One can thus establish a hierarchy of singularities and it is necessary to build a battery of theoretical concepts to analyze all the possibilities.

The interest of singularities — it was one of Thom's great ideas — is that if all the local singularities of an object are known, then the object can be globally known qualitatively. The singularities concentrate the qualitative information on the objects. I give an example, introduced by Moebius in the 19th century, but known for centuries by sculptors. A good way of understanding a three-dimensional form is to cut it out in two-dimensional slices and to consider these successive slices. Take a torus (considered vertically) and cut it out in horizontal slices of increasing height (see fig. 1). If the cutting plane is too low, you do not meet the torus. At a certain time you meet the torus at a first singular point (a minimum). When you still go up, you get circles. You still go up and you meet another singular point (a saddle): the section has the form of an 8. The following level lines contain two circles, then again one 8, then only circle, and finally you reach the point at the apex of the torus (a maximum).

Conversely, if by cutting out a surface in slices you meet four singular points of this type (a minimum, two saddles, a maximum), then the surface is topologically a torus. Topologically these four singular points (with their type) characterize a closed surface with a central hole. The fact that the list of the singular points with their type define the object topologically is called Morse theorem.

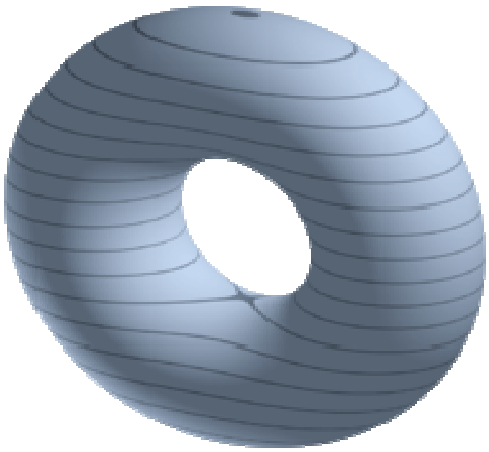


Figure 1. A torus with its level lines. Morse theorem
(source: http://fr.wikipedia.org/wiki/Théorie_de_Morse)

The classes of structures you can analyze with such methods are of a great diversity. You can look for example at what are called differentiable manifolds which generalize the concept of surface; you can also look at maps between spaces. In all these cases, you consider classes of objects and you want to study their possible non-trivial local properties.

The levels of structure can be very different from one another. You can consider topological objects (a very low level of structure but where the concept of continuity has a meaning nevertheless); or objects having more rigid properties. For example, if you take an orange peel and you try to crush it on a plane, as it is not elastic, it tears. This is a metric property: at the metric level, the sphere has curvature whereas the plane does not have any. Thom focused on the level of structure known as "differentiable", which is intermediate between topological and metric levels and means that you can take as many derivatives as you want of the functions describing the objects.

Q: What are Thom's contributions in linguistics?

To understand Thom' contribution in the fields of semiotics and linguistics, it is necessary to come back to the specific notion of structure in linguistics, which concerns the mereological problem: how totalities can be organized with constituents, relations and transformation rules between constituents, and show an organization which is more than the sum of their constituents.

There are many fields where mereological structures can be encountered: grammatical rules and syntax (whatever the theories), but also, in psychology, visual perception spatial objects linked by spatial relations; in biology, the constituent structure of organs; in chemistry, the molecules where atoms are linked by their valence electrons, etc. These structures have been pinpointed for a long time, but, until recently, the adequate mathematics to model them in biology and in linguistics was completely missing (in chemistry it is only with quantum mechanics that they could be modelled).

This is why the issue concerning structures refers to formalization and modelling. In linguistics, the mathematical models used are multiple but generally rest on formal tools, that is algebraic, combinatorial and logical tools (Chomsky, Shaumyan, Montague, etc). In other fields, like visual perception and biological morphogenesis, structures are interpreted in a much more geometrical, topological and dynamical manner, as organized forms and Gestalts.

The concept of structure is no longer algebraic and logico-combinatorial but morphological and dynamical, "morphodynamical" as I like to say.

The problem of a topological and morphodynamical mathematical theory for forms, primarily in biology and perception, is fundamental and extremely old. If we look back in history, there was a rather good theory in Aristote (cf. homeomeres and anhomeomeres in *The Parts of the Animals*) but it was completely eliminated by modern Galileo-Newtonian physics.

As André Robinet brilliantly showed, Leibniz was obsessed all his life by the antinomy thus created: one needs neo-Aristotelian concepts to work out a theory of form, but those seem to be incompatible with mechanist physics[7]. To overcome that antinomy, Kant had to write his third Critique, *The Critique of Judgment.* After him, many philosophers and scientists raised these issues. But these remained open until the 1960-70s when one suddenly saw flowering several radically new theoretical proposals: Thom and Zeeman with catastrophe theory, Ilya Prigogine with dissipative structures[8], Hermann Haken with synergetics[9], Henri Atlan with self-organization[10].

It is Thom who introduced the deepest mathematical tools. The only precedent had been, about fifteen years before, that of Turing who, just before his death, had been interested in morphogenesis and had introduced the first models explaining the emergence of forms and patterns in biochemical substrates using reaction-diffusion equations[11].

In the late 1960s, one thus started to have an idea of how the old problem of a theory of forms could be apprehended. Thom then introduced, in a very radical (and very controversial) way, the assumption that these morphodynamical tools could be transferred from biological morphogenesis to structural linguistics and semiotics. As a result, he found himself at the very heart of a linguistic debate which had a rich history: that of the opposition between gestaltic views (Guillaume, Tesnière, etc.) and formal views (Chomsky, etc.). Some linguists, like Hansjakob Seiler and Bernard Pottier, were enthusiastic. Others, like the Chomskyans, were more careful, even hostile.

I had the privilege to take part in the historical (and polemic) meeting between Jean Piaget and Noam Chomsky organized in 1975 at the Center of Royaumont by Massimo Piattelli-Palmarini, where I presented the principal differences between Chomsky and Thom[12].

[7] See André Robinet (1986) *Architectonique disjonctive, Automates systémiques et Idéalité transcendantale dans l'oeuvre de G. W. Leibniz*, Paris, Vrin. See also J. Petitot (1999). "Le troisième labyrinthe: dynamique des formes et architectonique disjonctive", *L'actualité de Leibniz: les deux labyrinthes* (D. Berlioz, F. Nef eds), *Studia Leibnitiana Supplementa*, 34, 617-632, Stuttgart, Franz Steiner.

[8] See for example Ilya Prigogine, Isabelle Stengers (1979) *La Nouvelle Alliance. Métamorphose de la science*, Paris, Gallimard.

[9] H. Haken (1981) *The Science of Structure: Synergetics* (Van Nostrand Reinhold).

[10] H. Atlan (1972/1992) *L'Organisation biologique et la Théorie de l'information*, Hermann,

[11] A. Turing (1952), "The chemical basis of morphogenesis", *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, Vol. 237, No. 641, pp. 37-72. See also J. Petitot (2013) "Complexity and self-organization in Turing", *The Legacy of A.M. Turing*, (E. Agazzi, ed.), Franco Angeli, Milano, 149-182. ArXiv: http://arxiv.org/abs/1502.05328v1. A. Lesne, P. Bourgine (eds.) (2006). *Morphogenèse. L'origine des formes*. Belin, Paris. Murray J.D. (2005) *Mathematical Biology*, Springer, New York.

[12] Petitot, J. (1979) "Hypothèse Localiste et Théorie des Catastrophes. Note sur le Débat", *Théories du Langage, Théories de l'Apprentissage, le Débat Chomsky / Piaget*, (M. Piattelli-Palmarini, ed.), 516-524, Le Seuil, Paris.

Q. How did Thom become interested in linguistics?

Thom much admired Lucien Tesnière. He deeply regretted that he did not meet this great linguist in Strasbourg when he was a young researcher working there with Henri Cartan between 1947 and 1951. His interest focused on the way Tesnière conceived the dependence relations between the constituents of a sentence (still the mereological problem!) and developed an almost narratologic idea of the sentence as a "scene" making actants interact: "The verbal node [...] expresses a small drama by itself. "[13]

Thom was philosophically a realist in linguistics. He estimated that, below the great variability and complexity of the morphosyntaxic surface structures, the universals of language result from evolution and are rooted in the cognitive abilities of primates, in particular in perception and action. Consequently, he tackled the linguistic problems from the point of view of the biological evolution of cognitive structures.

Q. Which other researchers besides you were interested in Thom's work on linguistics?

Among the linguists and semioticians who very early were deeply interested in Thom, one can quote, besides masters like Jakobson, Seiler and Pottier[14], two researchers of my generation: Wolfgang Wildgen[15] of the University of Bremen in Germany and Per Aage Brandt[16] of the University of Aarhus in Denmark. Their work developed, like mine, in the 1970-80s.

Then, in a completely independent way, without any reference to the European debate, something relatively similar happened in the United States in the 1980-90s with the emergence of West Coast cognitive linguistics: Charles Fillmore and George Lakoff at Berkeley, Len Talmy at Berkeley then at Buffalo, Ron Langacker at San Diego (where Gilles Fauconnier was, too). These linguists developed approaches which, on the one hand, were very structural (although with few references to European structuralism) and, on the other hand, explicitly supported the same theses on the evolutionary origin of language in relation to perception and action. At the time of the conference on Tesnière organized in 1992 in Rouen by Françoise Madray-Lesigne (Tesnière was born in 1893 in Mount-Saint-Aignan close to Rouen), Langacker made an emphatic praise of Tesnière by regarding him as one of his precursors[17] .

In addition, connectionist models of neural networks, opposed to formal models, were developing powerfully and rapidly. One of the linguists more in sight in that field was Paul

[13] L. Tesnière (1959), *Éléments de syntaxe structurale*, Klincksieck, Paris, 1959 (2ème éd. 1988), 48, 1. Voir aussi J. Petitot (1985) *Morphogenèse du Sens. Pour un Schématisme de la Structure*, PUF, Paris.

[14] See for example B. Pottier (2000) *Représentations mentales et catégorisations linguistiques*, Paris, Louvain, Peeters.

[15] cf. W. Wildgen (1982) *Catastrophe Theoretic Semantics. An Elaboration and Application of René Thom's Theory*, Benjamins, Amsterdam, and (1999) *De la grammaire au discours. Une approche morphodynamique*, Peter Lang, Bern.

[16] cf. Per Aage Brandt (1994) *Dynamiques du sens*, Aarhus University Press, Aarhus, and (1995) *Morphologies of Meaning*, Aarhus University Press, Aarhus.

[17] cf. Langacker, R.W. (1995) "Structural Syntax: The View from Cognitive Grammar ", *Lucien Tesnière aujourd'hui*, (F. Madray-Lesigne and J. Richard-Zappella, eds.), Actes du Colloque international CNRS, Université de Rouen 16-18 novembre 1992, Louvain/Paris, 13-39.

Smolensky whose models raised a violent debate with Jerry Fodor and Zenon Pylyshyn. I took part in that debate[18].

Wildgen, Brandt and myself made contact with these new trends. We organized meetings, for example in San Marino, at Umberto Eco and Patrizia Violi's "International Center for Semiotic and Cognitive Studies", a conference with Len Talmy, and also two important conferences at the CREA on the issue of constituent structures in connectionist models. These conferences were relayed by another one, organized this time at Bloomington by Tim van Gelder and Bob Port, entitled "Mind as Motion"[19].

Q. How were you informed of the works existing in the United States?

I was much interested in Fillmore — Case linguistics, with Tesnière's structural syntax, is the closest to Thom's views —, Langacker, Jackendoff, Lakoff. But the one that made contact with the most interesting linguist for us, namely Len Talmy, was Per Aage Brandt. Len achieved a splendid work on linguistic Gestalt while showing empirically, on a large corpus of data, the existence of very close connections between syntax (more precisely deep "syncategorematic" structures) perception and action. He went much further than case markers like prepositions[20].

Q. Which modelling for linguistics?

Once the empirical regularities are described, one asks how to model them. I regarded the works by Fillmore, Langacker, Talmy, etc. as well supported results; I trusted them, and what interested me was to see how the syntactico-semantic structures they had identified could be modelled.

For this purpose, I applied a general methodological principle. It is not because the structures under scrutiny are of a linguistic nature that the good tool, *a priori*, is formal languages. In sciences, one should not make any assumption on the fact that the mathematical tools should be of the same order as the objects. I am not anti-formalist *a priori*. I am ready to admit that formal models can prove to be the best in some cases. But I do not see any *a priori* reason that formal languages should constitute good tools to understand syntactic structures of natural languages, no more than to understand perceptive, biological or molecular mereological structures. My methodological principle is: the structures of natural languages are natural phenomena (I underline "natural") and, as in other sciences, it is necessary to invent (I underline "to invent"), starting from appropriate bases, the suitable mathematical tools to model them.

These appropriate bases are two-fold. On the one hand, they come from properly linguistic studies and on the other hand from other disciplines like cognitive sciences and

[18] See for example P. Smolensky (1988). "On the Proper Treatment of Connectionism", *The Behavioral and Brain Sciences*, 11, (1988), 1-74. J. Fodor, Z. Pylyshyn (1988) "Connectionism and cognitive architecture: a critical analysis", *Cognition*, 28, 1-2 (1988) 3-71. J. Petitot (2011) *Cognitive Morphodynamics. Dynamical Morphological Models of Constituency in Perception and Syntax* (with R. Doursat), Peter Lang, Bern.

[19] T. van Gelder, R. Port (eds.) 1995, *Mind as Motion*, Cambridge, MIT Press.

[20] Len Talmy (2000). *Toward a Cognitive Semantics*, Vol. I: *Concept Structuring Systems*, Vol. II: *Typology and Process in Concept Structuring*, Cambridge, MA, MIT Press, 2000.

neurosciences. Of course, no need to make brain imagery to explain the conditional mood in French, but one must use neurocognitive results on universal sensorimotor schemes of interaction between actants to understand verbal valence.

III. Phonetics, phonology and catastrophe theory

Q. You were interested in phonetics.

In certain cases, non-formal mathematical models of a topological-geometrical-dynamical type have proved to be rather good. The first example which convinced me of the validity of morphodynamical models in the field of language was phonetics. As you know, structuralism comes mainly from the phonological work of the Moscow and Prague Circles, and, when I began to work with Thom, I already knew structural phonology a little and the remarkable results that Jakobson transfered to general structuralism, in particular in collaboration with Claude Lévi-Strauss. I thus tried to test Thom's models on it and I discovered (I consider that it was my first scientific "discovery") that they were completely adapted to phonology.

It happens that at the EHESS there was (and there still is) a very good laboratory of cognitive psychology[21] some members of which worked in phonetics. Following their advice, I read many things in this field and I noted that Thom's models were not only relevant but that they were quantitatively and qualitatively exact.

In phonetic perception, several levels should be distinguished: the acoustic level, the peripheral level of sensory transduction, the perceptive level and the linguistic level. At one end of the chain, one can make a lot of acoustico-physical experiments and at the other end one has at one's disposal a very important linguistic corpus of thousands of languages.

One characteristic of phonetic perception is what is called its "categorical" character. What does that means? When one makes a sonogram one can identify the "formants" of the sounds produced by the vocal tract. Sounds produced by vocal cords are very rich in harmonics, and the articulatory controls control the shape of the resonators of the vocal tract. Each of these resonators (mouth, nose …) amplify or damp specific harmonics. In other words, the amplitudes of the harmonics are modulated by a continuous curve having strongly marked peaks. These resonance peaks select frequency bands which are called formants. Vowels are stationary sounds having characteristic formants, and consonants are transient sounds carrying out transitions between formants and possibly introducing turbulence (plosives, fricatives).

When you look at the equations, you observe that the formants correspond to the maxima of what is called "the transfer function" (the output/input ratio) of the vocal tract. In fact this function H is the inverse of a function G and the maxima of H correspond to the minima of G.

One can simplify the problem by preserving only a few resonators, for example three: the front cavity (mouth), the back cavity (pharynx) and the nasal cavity. Each cavity is described by a tube (with length and diameter) and constrictions of the vocal tract are described by small intermediate tubes. One knows how to explicitly compute the way in which the formants depend on these articulatory parameters. Personnally, I used as a guide the classic *Preliminaries to Speech Analysis* by Jakobson, Fant and Halle[22]. From this audio-

[21] The LSCP (Laboratoire de Sciences Cognitives et Psycholinguistique).

[22] Jakobson, R., Fant, G. & Halle, M. (1952) *Preliminaries to Speech Analysis. The distinctive features and their correlates*, MIT Technical Report.

acoustic base, structuralist works, in particular Jakobson's, show how phonological (linguistically relevant) distinctive features can be recovered using a qualitative description of the formant configurations. For example if one considers the universal vocalic triangle /*a*/, /*i*/, /*u*/ in simple models with two formants:

/*a*/ corresponds to close formants of medium frequency (feature "compact"),
/*i*/ corresponds to well separated formants (feature "diffuse") with predominance of the "acute" formant (high frequencies),
/*u*/ corresponds to well separated formants (feature "diffuse") with predominance of the "bass" formant (low frequencies).

If more detailed models are used, one can still qualitatively describe phonological distinctive features in this way by using not the true formants quantitatively defined, but "formantial masses" as Ludmilla Chistovich proposed a long time ago.

Then, I looked at the explicit formulas connecting the formants to articulatory controls and I discovered that, for the models with tubes, the function *G* exactly is an unfolding of singularity in Thom's sense and that the formants and their configurations are consequently describable in terms of catastrophes: there are abrupt — discontinuous — changes in formantial masses according to continuous changes in articulatory controls[23].

Let us be a little more technical. The transfer function is a function *H(s)* of a complex variable $s = \sigma + i\omega$ where $\omega/2\pi$ is the frequency and σ a damping factor. The restriction of *H*(*s*) to the imaginary axis ω gives the modulation of the harmonics frequencies. Let us consider the model with one resonator (i.e. two tubes and one formant) of figure 2.

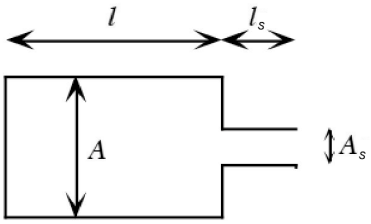


Figure 2. Model with two tubes

One obtains (for conveniently chosen values of *l*, l_s and A_s) the already complex formula:

$$H(s) = \frac{1}{LC\,s^2 + (RC + GL)\,s + GR + 1} = \frac{{\omega_0}^2}{(s - s_1)(s - s_2)}$$

with

[23] See J. Petitot (1985) *Les Catastrophes de la Parole. De Roman Jakobson à René Thom*, Maloine, Paris; and (1997) "Modèles morphodynamiques de catégorisations phonétiques", *The Roman Jakobson Centennial Symposium* (P.A. Brandt, F. Gregersen eds), *Acta Linguistica Hafniensia*,29,239-269. http://jeanpetitot.com/ArticlesPDF/Petitot_Jakobson.pdf

$$L = \rho l_s / A$$
$$C = lA / \rho c^2$$
$$R = \frac{S}{A^2}\sqrt{\frac{\omega\rho\mu}{2}}(l + l_s)$$
$$G = S\frac{\eta - 1}{\rho c^2}\sqrt{\frac{\lambda\omega}{2c_p\rho}}(l + l_s)$$

,

where A = section of the open tube, S = circumference of diameter A, ρ = density of the air, c = speed of sound, μ = coefficient of viscosity, λ = coefficient of conduction of the heat, η = adiabatic constant, c_p = specific heat of air under constant pressure. The poles of $H(s)$ are

$$s_1 = -\frac{1}{2}\left(\frac{R}{L} + \frac{G}{C}\right) + i\sqrt{\frac{GR+1}{LC} - \frac{1}{4}\left(\frac{R}{L} + \frac{G}{C}\right)^2}$$
$$s_1 = \sigma_1 + i\omega_1$$
$$s_2 = \bar{s}_1 = \sigma_1 - i\omega_1$$
$$\omega_1 = \sqrt{{\omega_{01}}^2 - {\sigma_1}^2}\ , \quad {\omega_{01}}^2 = \frac{GR+1}{LC}$$

Only s_1 is relevant because its imaginary part is > 0 and frequency must be > 0.

For a model with two resonators (four tubes and two formants), one obtains the formula:

$$H(s) = \frac{\omega_1\,\omega_2}{s^4 + as^3 + bs^2 + cs + d}$$

with

$$\omega_1 = \frac{1}{\sqrt{L_1 C_1}}\ ,\ \omega_2 = \frac{1}{\sqrt{L_2 C_2}}$$

$$a = \frac{R_1}{L_1} + \frac{R_2}{L_2} + \frac{G_1}{C_1} + \frac{G_2}{C_2}$$

$$b = \frac{1 + R_1G_1}{L_1 C_1} + \frac{1 + R_2G_2}{L_2 C_2} + \frac{R_1G_2}{L_1 C_2} + \frac{R_2G_1}{L_2 C_1} + \frac{G_1G_2}{C_1C_2} + \frac{R_1R_2}{L_1 L_2} + \frac{1}{L_2 C_1}$$

$$c = \frac{\left(R_1(1 + R_2G_2)\right)}{L_1L_2\,C_2} + \frac{\left(G_2(1 + R_1G_1)\right)}{L_1C_1\,C_2} + \frac{R_1 + R_2 + R_1R_2G_1}{L_1L_2\,C_1} + \frac{G_1 + G_2 + G_1G_2R_2}{L_2C_1\,C_2}$$

$$d = \frac{\left(1 + R_1G_1 + R_2G_2 + R_1G_2 + R_1R_2G_1G_2\right)}{L_1C_1L_2\,C_2}$$

Figure 3 shows the graph of the logarithm of the module of $H(s)$ and the damping of the formants.

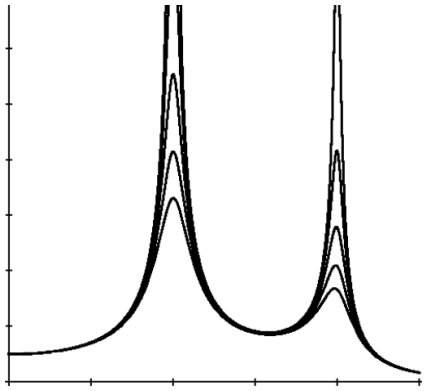


Figure 3. Damping of the two formants.

The key point is that, for n formants, (*i*) the denominator of the transfer function $H(s)$ is the universal unfolding of the singularity known as A^{2n}

$$A^{2n}(s) = s^{2n} + a_{2n-1}\, s^{2n-1} + \quad + a_1\, s + a_0$$

and (*ii*) the coefficients of this unfolding are complicated functions of the $2n$ articulatory controls.

To move from formants to formantial masses, and thus from quantitative to qualitative, one introduces an "auditive transformation" which merges the sufficiently close formants. Then one really obtains models in the sense of Thom.

In short, one can, thanks to Thom's models, explicitly move from the acoustic level (physical) to the auditive level (sensorial) then to the phonological level (linguistic), the key being the interpretation of formantial masses in terms of unfoldings of singularities parameterized by articulatory controls.

Let us address now the issue of categorical perception[24]. In addition to articulatory controls determining the shape of the vocal tract, there exist other acoustic cues that one can vary in a continuous manner, e.g. voicing (VOT: voice onset time) that measures the moment of excitation of the fundamental harmonic. In short, while one can vary many parameters continuously, perception does not vary continuously. It is the fundamental reason why some sounds can be the substrate of a phonological code. For example, you can vary voicing in order to move from [b] (voiced labial) to [p] (not-voiced labial). But at the perceptive level, on the other hand, you perceive only allophones of /*b*/ or /*p*/ and no intermediate state.

To explain this remarkable phenomenon, psychologists distinguish two fundamental mechanisms. On the one hand, discrimination: can I discriminate two close [b]; and on the other hand, identification: do I identify a /*b*/ or a /*p*/, i.e. a sound as an allophone of a phoneme or of another one.

For colours, discrimination corresponds to shades and identification corresponds to colour names. The perception of colours is "continuous" in the sense that shade discrimination depends very little on colour identification: one perceives gradual shades

[24] Among a rich bibliography, three references were important for me: Liberman, A. M., Cooper, F. S., Shankweiler, D. P., Studdert-Kennedy, M., (1967) Perception of the Speech Code, *Psychological Review*, 74, 6, 431-461. Stevens, K., (1972) The Quantal Nature of Speech, *Human Communication, a Unified View* (P. B. Denes, E. E. David Jr. eds.). Malmberg, B., (1974) *Manuel de phonétique générale*, Paris, Picard.

independently of the existence of the categories of colours. In categorical perception, the situation is quite different: discrimination degenerates inside the categories; as one says, it is subordinated to identification: one discriminates two close sounds only if they are identified as different. One is unable to discriminate two close [b] sounds identified as allophones of /*b*/, but on the other hand one is able to discriminate two close intermediate sounds if one is identified as an allophone of /*b*/ and the other as an allophone of /*p*/.

It is a little as in geography: there are areas delimited by boundaries (the domains of the parameter space corresponding to a single phoneme), inside an area the various positions (allophones) have an equivalent type (they are tokens of the same phoneme: no intra categorial discrimination), but at the boundary crossing, the type (the corresponding phoneme) changes abruptly. In categorical perception, there exist thresholds between categories which are induced by perception itself. That is due to the fact that percepts vary in an extremely non-linear way compared to their audio-acoustic and articulatory controls. Kenneth Stevens well studied this phenomenon in his article quoted above "On the quantal nature of speech".

Categorical perception is a fundamental property of phonetic perception and, I repeat, explains why and how some sounds can become the substrate of a code.

Catastrophe theory is particularly well adapted to the modelling of categorical perception because its general model rests on the concept of "bifurcation". A bifurcation occurs in a system when a small change of a continuous control produces a qualitative jump of the internal state of the system, in other words when a small variation of causes involves great differences on effects. It is precisely what occurs with categorical perception when, for example, a small articulatory change qualitatively moves the configuration of formantial masses from a single formantial mass to two formantial masses ("compact/ diffuse" opposition).

I thus showed that the catastrophes related to audio-acoustic equations match the phonological structures observed in languages. For phonetics, Thom's models are thus valid models. Jean-Luc Schwartz and the Grenoble group, among them Christian Abry and Louis-Jean Boë, went more thoroughly into them[25]. In particular they identified the "auditive transform" as a mechanism of "large scale spectral integration".

To summarize, catastrophe models help understanding in a detailed way the link between audio-acoustics, psycho-physics, perception and structural phonology.

Q. What exactly is a catastrophe model?

A catastrophe model starts with a system which has internal states. For instance, in the case of phonetic perception, there are neuronal states corresponding to percepts. In the case of a chimical element such as water, the thermodynamical states are called "phases": solid, liquid, gas. These internal states are attractors of the internal dynamics of the system and the transient states, induced by the inputs of the system, are rapidly stabilizing toward them: for instance, an acoustic input turns into a perceptive state (after having gone through the external ear, the cochlea, the auditory cortex). Moreover, the system is controlled by external parameters (articulatory parameters, acoustic cues, temperature, pressure…). When these controls change, the inner states change in turn, and there is two possible outcomes. Either a

[25] See for example Abry, C., Boë, L.-J., Schwartz, J.-L. (1989). Plateaux, catastrophes and the structuring of vowel systems. *Journal of Phonetics* 17, 47-54. Schwartz, J.-L., Boë, L.-J., Vallée, N., Abry C. (1997). The dispersion-focalization theory of vowel systems. *Journal of Phonetics* 25, 255-286.

small change in the controls is without consequences and does not change the inner state qualitatively (e.g. the sound is still perceived as a /*b*/, the water temperature shifts from 50°C to 51°C), or a small change in the controls changes the qualitative type of the inner state (the sound is now perceived as a /*p*/, the water temperature shifts from 99°C to 100°C and the water starts boiling). Such qualitative changes are called “bifurcations”. In thermodynamics they are called “phase transitions”.

Q. Can a shift from /*b*/ to /*p*/ be predicted, in the sense that it follows some constraints?

There are strong differences across languages. But one can assume a universal innate “initial state” for new born humans. During language acquisition, some thresholds move, others split, others disappear. For instance, in the case of the Japanese language, the threshold between [r] and [l] disappears and [r] and [l] become two allophones of a single phoneme. Young Japanese are able to discriminate between [r] and [l] but, while learning the language, these discriminations disappear due to the categorical property of perception.

Thus, maybe there is an initial phonological “geography” that evolves with the learning of a specific phonological system. All the phonological systems categorize the same space of sounds defined by anatomically possible articulatory controls and harmonics. The question is to know whether the categorisations of actual languages are efficient and whether they reach an optimum of the quantity of information conveyed by the phonological code.

Q. Is there a definition of what is an optimum vocalic system for communication?

This is a fascinating question. We know of a great number of phonological systems and they may be grouped into classes. Numerous models have been proposed. The problem is the following: within the space of the possible sounds, defined by universal anatomic constraints, we have to find the best categorisation into sub-regions (the phonemes). There are several strategies in order to solve the problem of the optimisation of the categorisation and there are several studies showing how these strategies are related to each other. All the phonological systems are based on the universal vocalic triangle /*a*/, /*i*/, /*u*/, and can be described as a progressive refinement complexifying that triangle, leading eventually to the most complicated vocalic systems, such as that of French.

Q. We mentioned “basins of attraction”. Could you explain that notion more thoroughly?

When you have a dynamics defined on a given space *M*, all the points *x* belonging to *M* have a trajectory $\gamma(x)$ and you can consider the asymptotic behaviour of that trajectory. Generally, $\gamma(x)$ is attracted by an attractor *A* which is a sub-set topologically closed and dynamically invariant, minimal for these two properties, and which attracts all the trajectories coming from points in its neighbourhood. All the points *x* whose trajectory $\gamma(x)$ are asymptotically attracted by *A* constitute the basin of attraction *B*(*A*) of *A*. Thus the dynamics decomposes *M* in several basins of attraction separated by boundaries (some of them can be very complicated).

In the case of the internal dynamics of a system, every input puts the system into an initial state x and, generally, x is not on the boundary but inside a basin of attraction $B(A)$ and is therefore attracted by the attractor A. This means that the initial transient state x will be attracted by the internal state A. This projection of the input on an attractor models the process of "identification". In this type of models for categorisation, the basins of attraction are the categories and the attractors are the prototypes. An input induces an initial state that is associated with a prototype. In the phonetic domain, a sound is recognized and identified as an allophone of a given phoneme.

Some boundaries between basins of attraction are more complex than others. When an initial state is on a boundary separating two basins as a sort of ridge-line, it is possible to fall into one or the other of the basins. And, last but not least, the control space allows the modification of the basins of attraction and their boundaries.

That is why Thom proposed to distinguish two kinds of bifurcation: (*i*) internal bifurcations, when the system moves from a basin of attraction to another in the internal space M and (*ii*) external bifurcations, when the system is coerced into another attractor by the effects of the controls, for instance due to the fact that an attractor disappeared or that two attractors have merged. In practice, external bifurcations matter most: the systems are in general of the "slow/fast" kind, which means that the internal dynamic is fast, while the variation of the controls is slow and then it is possible to do as if the system were always in an internal stable state (on an attractor). It is called an "adiabaticity hypothesis". What chiefly matters then is the bifurcation of the attractors and not the fast internal transient trajectories.

Q. From a mathematical point of view, which kinds of mathematics are involved?

When building the mathematics for his models, Thom chose – for the spaces, the functions and the maps between spaces he needed – the level called differentiable, that is the level where the objects have locally almost everywhere well-defined derivatives, except sometimes in some singular points. This level is more constrained than the basic continuous level (he does not allow objects such as fractals). However, it is far less constrained than the algebraic or metric level. The differentiable objects are very "flexible".

Thom introduced two kinds of models: the elementary models, and the extended ones. In the first ones, the internal dynamics is the steepest descent of an energy potential function $f(x)$ defined in the inner space M: the system optimizes its state by minimizing its internal energy. The attractors (the internal states) are in that case the minima of $f(x)$: an initial state goes (according to the specificities of the system) either to the closest minimum or to the absolute minimum.

The control parameters allow the variation of the potential functions, and therefore the change of the minima and their height. A minimum may then disappear and the system will have to go into another minimum: these are bifurcations.

One of Thom great achievements has been the (difficult) proof of the classification theorem for elementary catastrophes. The main idea is the following: if you consider a potential function f where several minima, maxima or saddles (called "critical points") are merged in a single point x, f has an unstable singularity at x (according to a natural notion of stability). If you deform such a singularity through external small parameters w embedding f into families $f_w(x)$ defined in a small neighbourhood of x with $f_0(x) = f(x)$, it is possible to stabilize the singularity in many ways, partially or totally, through the dissociation of the critical points that have been merged. The key result is that, given such a singularity, there exists a universal

deformation, called “universal unfolding”, that gathers optimally all the possible stabilisations.

Figure 4 shows the catastrophe named “cusp”, that plays the key role in modelling the universal vocalic triangle. The unstable singularity x^4 merges two simple minima (non-degenerated minima, i.e. that are not composed by merging simpler critical points) and one simple maximum. The external space W of the universal unfolding is two dimensional. It is partitioned into three regions by a catastrophe set K, containing the two branches K_b of a cusp curve and the median half-line K_c. Along the branches K_b one simple minimum remains simple while the other simple minimum and the simple maximum merge into an inflection point: the K_b are lines of catastrophes of bifurcation. Along K_c, the two simple minima and the simple maximum remain simple but the two minima have now the same height: K_c is a line of catastrophes of conflict. Apart from K, $f_w(x)$ have either a simple minimum, either two simple minima separated by the simple maximum with one of the minima that dominates the other.

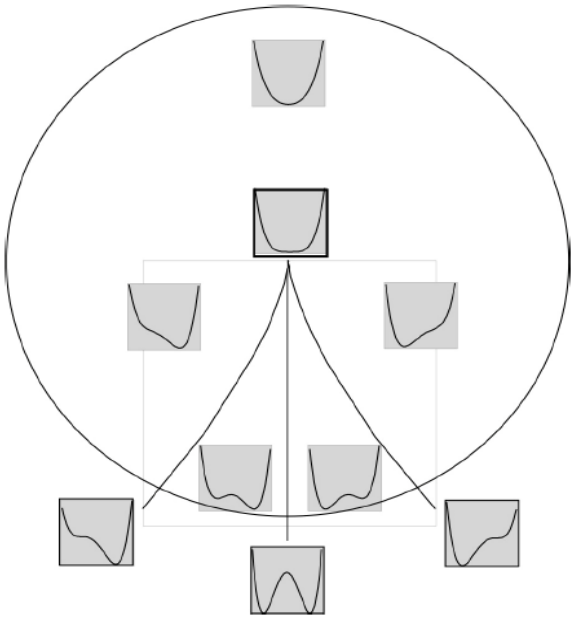


Figure 4. The universal unfolding of the “cusp” catastrophe

The classification theorem says that, whatever the system under scrutiny, if there are one or two internal dimensions and no more than four external controls w, and if the potential function $f(x)$ of the system has unstable singularities, and if its unfolding is structurally stable (that is the process stabilizing internal unstabilities is itself stable), then these singularities belong to a finite list: “cusp”, “swallowtail”, “butterfly” in the one dimensional case, elliptic, hyperbolic or parabolic “umbilic” in the two dimensional case.

From a methodological point of view, this result is very important because it exhausts the field of possibilities. It is as important as the theorem of classification for the platonic solids (the finite sub-groups of the group of the rotations) [26].

[26] Readers interested in the theory of mathematics may wish to read (in French) : Chenciner, A., (1980). “Singularités des fonctions différentiables”, *Encyclopædia Universalis*, Paris ; as well as the compilation I made in 1982 (in French): *Eléments de théorie des singularités*, http://jeanpetitot.com/ArticlesPDF/Petitot_Sing.pdf

III. Language and perception

Q. What are the consequences for linguistics?

As I said Thom was interested in a realist approach to language. For him, language had an evolutionary origin. The ability to describe perceptive scenes of the outer world communicate them to those that do not see them was, for him, a fundamental requisite constraining natural languages. It appears that a large part of these perceptive scenes are interactions between "actants"[27] (being either agents or objects), and the transformations of their spatial relations can be described by verbs (to go into a place, to seize something, to attack, to run away, etc.).

The basic assumption is that the structure of the sentences describing a perceptive scene with verbs and actants situated in space and time is a result of the evolution pressure and that there is an analogy between the constituent structure (mereology) of actantial syntax and the constituent structure of the perceptive scenes.

This assumption of a foundation of the actantial structures in the structures of perception and action has a long history. One of its components is the "localist hypothesis", which has been supported under various guises by linguists such as Anderson, Langacker or Talmy, and according to which the basic syntactic structures of elementary sentences categorize the generic interactions in space and time. Structural syntaxes like Tesnière's and case grammars like Fillmore's belong to the same paradigm. All these theories rely on an actantial theory using semantic roles defined by spatio-temporal schemas similar to schemas of perception and action and therefore rooted in cognitive evolution.

In this context, Thom's theorem is of primary importance: it is possible to classify the actantial spatio-temporal interactions thanks to the classification of elementary catastrophes. This theorem proves the existence of case universals. It is obviously a fundamental result.

These hypotheses have been controversial. Numerous linguists objected that language is independent from perception and that it is a cognitive faculty *sui generis*. Chomsky for instance argues for the notion of autonomy of syntax. Other linguists acknowledge the existence of links with perception but argue that the linguistic categories of perception cannot be extracted from the perception itself. And, in any case, these hypotheses are but of low interest for linguists describing actual natural languages since they pertain to a deep "proto-linguistic" level, far below from the morphosyntactic diversity of natural languages. However, these hypotheses have a great theoretical significance for building bridges between linguistics and cognitive neurosciences. They have also a great technological relevance, in order to build robots able to convert natural language instructions in terms of perceptual structures and motor programs.

Q. Could you elaborate upon the relation between localist hypothesis and catastrophe theory?

If one try to schematize perceptual scenes and actantial relations (schematization is a strong simplification focusing only on the essential forms), one encounters again structures that are derived from elementary catastrophe.

[27] We use the term "actant" analogue to what are called "semantic roles" in case grammars. It is a deeper concept than that of "actor" or "character".

Let's take objects distributed in space, that is to say static configurations of spatial actants. Let us add a temporal evolution that changes this configuration dynamically. These temporal evolutions generally lead the actants to interact. It is then possible to represent the actants through minima of a potential function (here is the schematization) so that to turn the interactions into bifurcations and so that it is possible to apply the models of elementary catastrophes. I explain this in details in *Cognitive Morphodynamics*[28]. A schema such as "to take an object" is the fact that there is an actant and an object which are initially disjoint and which, later, are conjoint. The verbal node lexicalized by the verb "to take" describes this interaction, which is a bifurcation derivable from the cusp catastrophe. As soon as early 1970s, Thom made the list of the "archetypal actantial graphs" that are derivable from elementary catastrophes[29].

Later on, Thom's archetypes have turned out to be great precursors of several cognitive models of language: Fillmore's frames, Langacker, Talmy, Lakoff's image-schemas, Haiman's "iconicity in syntax", Desclés' "cognitive archetypes", Shank and Abelson's scripts, etc. (for more details see *Cognitive Morphodynamics*[30]).

Using the fact that elementary verbal nodes grammaticalize bifurcations of actantial relations, you can build a theory of verbal valency. It was one of the results that mattered the most for Thom. All the linguists that have been interested in verbal valency know that there is a limit of 4 actants (the few controversial cases with 5 actants use indeed a double actant). Where does this limit come from ? Why could not we be able to create new semantic roles allowing for an increase of the valency ? According to Thom, it is one of the strongest evidences of the rooting of actantial syntax into perception and action. Indeed, perception and action take place in a 4-dimensional space-time, and archetypal actantial graphs derive from elementary catastrophes whose external space have 4 dimensions at most. This closed list of catastrophes, drawn from the classification theorem, puts a drastic limit on the complexity of the bifurcations and, then, on the verbal valency. We then observe that, in all archetypes, the valency has a limit of four[31]. According to Thom, this constraint comes from our outer world.

Of course, several linguists objected that, in most of the verbs denoting action, there is an agentivity, and that agents are generally intentional agents. However, numerous remarkable experiments have shown how strongly agentivity itself is deeply rooted in perception and action. As early as the 1940s, F. Heider and M. Simmel have shown that movements, even complex ones (such as movements including accelerations, decelerations, changes of direction, etc.) of simple forms (triangles, circles, rectangles of various sizes) were spontaneously described by way of intentional action verbs ("come in", "come out", "give", but also "hide", "escape", "hunt", "attack", "force", etc. [32]). Since these pioneering experiments, numerous works were devoted to such phenomena. Let us mention for instance J. Scholl and P. D. Tremoulet on the perception of causality and the animacy of objects; D.

[28] Peter Lang, Bern, 2011.

[29] See for instance "Topologie et linguistique", *Essays on topology and related topics*, A. Haefliger and R. Narasimhan (eds), Springer, 1970, 226-248. Reprinted in: *Modèles mathématiques de la morphogenèse*, Paris, 10-18 UGE, 1974.

[30] See for instance Fillmore, C., (1976) "Frame semantics and the nature of language", In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*. Volume 280, 20-32. Haiman, J., (ed.) (1985) *Iconicity in Syntax*, Amsterdam, J. Benjamins. J.-P. Desclés (1990) *Langages applicatifs, Langues naturelles et Cognition*, Paris, Hermès. Schank, R., Abelson, R.P. (1977) *Scripts, Plans, Goals and Understanding*, Hillsdale, Lawrence Erlbaum.

[31] With 4 as the number of dimensions d of space-time and with 4 as the maximum valency $v(d)$, we have $v(4) = 4$. But it is not the case that $v(d) = d$ generally.

[32] Heider, F., Simmel, M. (1944) "An experimental study of apparent behavior", *American Journal of Psychology*, 57 (1944) 243-259.

Premack on the perception of intentional movements by children; S. J. Blakemore and J. Decety on the comprehension of intentions; M. E. Zibetti on the fact that we interpret as if the movements we perceived were caused by intentional agents [33]. All these works try to unveil the evolutionary and cognitive roots of the tendency we have to interpret purely cinematical and dynamical motions as if they resulted from an intentional agentivity. All these authors showed that this tendency is automatic, non-conceptual, "hardwired" and "rooted in automatic visual processing". Their works offered a general confirmation of Thom's theses.

Q. How to build a model of the perceptive scene?

This is an old and interesting issue. In order to move from perception to language, it is necessary to describe the perceptive scenes very schematically in order to specify the relevant information that should be translated into elementary proto-sentences. But how to simplify the perceptive scenes in a bottom-up and data-driven way, and how to define the linguistically relevant information that lays inside?

For instance, let us consider a preposition as "across" (this example is drawn from my book *Cognitive Morphodynamics*). This preposition can be applied to a huge diversity of perceptive scenes. "Across" can be applied to the crossing of a street, a lake, a field, a country, to walking haphazardly into a forest, etc. How would it be possible to define the geometric and topological invariant content of "across"? Obviously, that invariant pertains to the notion of "transversality", but how could we design algorithms extracting such a schema from a real and complex scene? The problem is to find good tools for simplifying the scene, tools strong enough for skeletonizing it, but also able to preserve the relevant information (" transversality "). In *Cognitive morphodynamics*, I showed, together with René Doursat, that some morphological algorithms implemented in neural networks can do the job[34].

Figure 5 schematizes the clause "zigzagging across the woods": (*a*) is the input image; it contains two objects: the path and the wood. (*b*) and (*c*) result from a first preprocessing using morphological algorithms of dilation and skeletonization. (*d*) and (*e*) result from a second preprocessing; in (*e*) the intersection of skeletons *sk*(*A*) and *sk*(*B*) allows one to extract the invariant schema of "transversality".

[33] See for example Scholl, B.J., Tremoulet, P.D. (2000) "Perceptual causality and animacy", *Trends in Cognitive Science* 4(8), (2000), 299-309. Blakemore, S. J., Decety, J. (2001) "From the perception of the action to the understanding of intention", *Nature Reviews Neuroscience* 2, (2001), 561-567. Zibetti, E., Tijus, C. (2003) "Perceiving Action from Static Images: the Role of Spatial Context", *CONTEXT* (2003) 397-410.

[34] These algorithms are discrete variants of dynamic processes of diffusion and skeletonization often used in morphodymics. They are computationally very efficient. They have been developed by G. Matheron and J. Serra and their colleagues. See for instance: Serra, J. (1982). *Image Analysis and Mathematical Morphology*, New York, Academic Press, 1982.

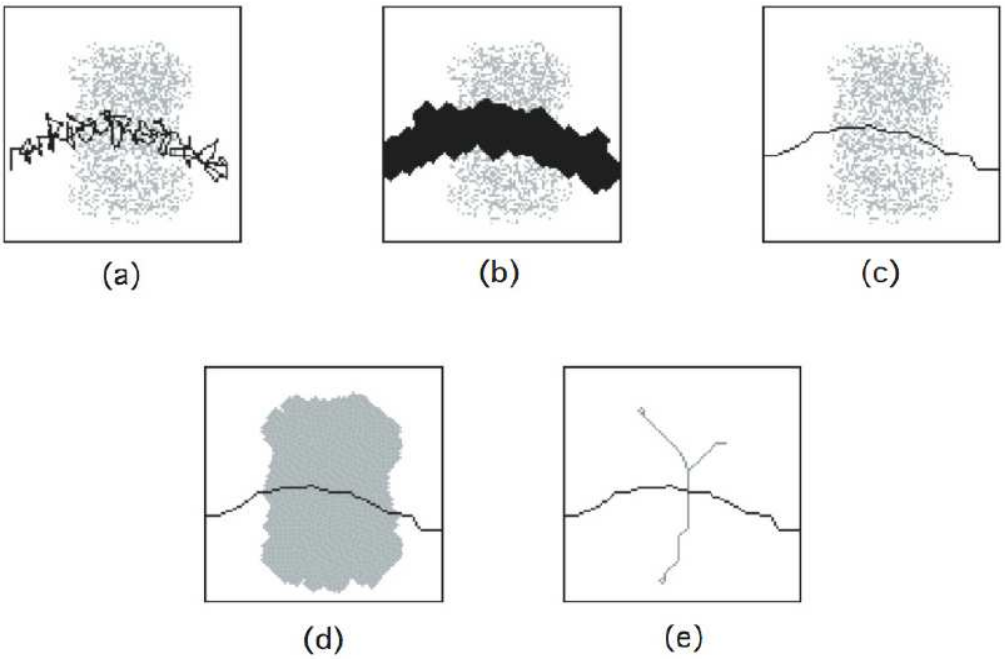


Figure 5. The extraction of the invariant content “transversality” in a pixelized image described using the preposition “across”.

I stress the fact that these morphological analyses of images are not obvious at all and that our capacity to easily and correctly apply prepositions to visual scenes is far from being understood. This issue is far most complex than that, already quite difficult, of hand writing recognition.

IV. Modelling and simulation

Q. What is the relation between modelling and simulation?

Models are mathematical, but do not belong to reality. I am not a realist concerning mathematics. Let us consider classical physics. Newton’s equation is perfect, but planets do not do differential calculus. They move, but they do not solve equations. However, Newton’s equation allows for computing (either explicitly or only numerically) some solutions that simulate perfectly the observed motions. The same holds for all models. We start with collecting a great corpus of empirical data and then we try to find good models able to generate a virtual reality that simulates the empirical reality.

If the morphodynamical models based on the hypothesis of the rooting of language in perception are correct, then it would be right to try to understand their neuronal implementation. It is not easy at all. Stephen Kosslyn, a well known neurophysiologist of vision, has studied with current methods of brain imaging the neural activity during the use of prepositions. He showed that there exist two systems for the processing of spatial relations: one is a continuous quantitative processing (*A* is more or less above *B*), and the other is a categorical discontinuous processing (*A* is above or beside *B*). Moreover, the two neural processings are lateralized: the continuous one takes place in the right hemisphere, while the

categorical one takes place in the left hemisphere[35]. Hence, if one wants to know exactly how the brain deals with prepositions, he has to go deeply into the analysis and modelling of the link between perceptive structures and linguistic categorization.

Interested readers will find more information in *Cognitive Morphodynamics* as well as in my 2008 book *Neurogéométrie de la vision. Modèles mathématiques et physiques des architectures fonctionnelles*[36] in which I deal with the neural implementation of basic properties of perception (which are already very difficult to understand even though they remain very far from the complexity of language).

Q. Is simulation a form of explanation?

It depends on the structure of the models on which the simulation is based. "It works!" is not by itself an explanation since it can pertain to the mere fine-tuning of ad hoc parameters. Models are explanatory when they arise from general and strong hypotheses while being able at the same time to generate good simulations. It is the case with Newton's equation, which results from general physical principles; it is the case with the elementary catastrophes which result from general principles of structural stability and from the dimensions of space-time.

Q. Numerous linguistics phenomena are quantitatively characterized by a Zipfian distribution. Is there any relation between this characteristic and the modelling proposed by catastrophe theory?

I have never worked in the field of statistical linguistics. Regarding Zipf's law in particular, I haven't worked on this subject, although the CAMS did work a lot on it[37].

However, one can't ignore that statistics are a good way for approaching regularities and that, during acquisition, children are learning rules in a statistical way: they extract linguistic rules by generalizing over a finite set of examples.

There are very interesting connectionist models (those by Jeffrey Elman seem to me to be the most interesting) that model how the syntagmatic statistical regularities induce semantic paradigms.[38]. You consider a small corpus containing various classes of nouns (animate agent, non animate object, etc.) and various classes of verbs (*to eat, to read*…). Then you do supervised learning with a neural network: you give a word as input to the network, you ask it to add one more word, and you correct it if it outputs an incoherent sentence. At the beginning, the network produces outputs that haven't any coherence, neither syntactically nor semantically. The corrections you pointed-out allow it to change its internal structure (i.e. to change the weight of its hidden layers) by retro-propagating the errors. When learning is

[35] Kosslyn, S.M. (2006). "You can play 20 questions with nature and win: Categorical versus coordinate spatial relations as a case study", *Neuropsychologia*, 44 (2006) 1519-1523. See also Kemmerer, D. (2007) "A Neuroscientific Perspective on the Linguistic Encoding of Categorical Spatial Relations", *Language, Cognition and Space*, (V. Evans et P. Chilton eds), *Advances in Cognitive Linguistics*, London, Equinox Publishing Co.

36 Les Editions de l'Ecole Polytechnique, Distribution Ellipses, Paris.

37 Cf. Micheline Petruszewycz, "L'histoire de la loi d'Estoup-Zipf: documents", *Mathématiques et sciences humaines*, 44 (1973) 41-56.

[38] Cf. Elman, J. (1989). Representation and Structure in Connectionist Models, *Cognitive Models of Speech Processing*, (G. T. M. Altmann, ed.), Cambridge, MA, MIT Press, 1989, 345-382.

done, the network does not make errors anymore. Then, you look at its hidden layer, and you see that it has built paradigms (animate agents, inanimate objects, state verbs, transitive and intransitive action verbs, etc.). "Paradigm" here means that the words are grouped into clusters. In other words, in order to produce correct syntagmatic sentences, the network has built some semantic rules.

Q. In the field of complex systems, what are the differences today between dynamical models and connectionnist models?

The difference between (morpho-)dynamical models and connectionist models is the following: connectionist models do have internal dynamics and, hence, attractors. They are made of atomic units (the formal neurons) linked by inhibitory/excitatory connections having synaptic weight. Each unit influences the units to which it is connected, which produces a global internal dynamics of the network.

The main interest of these connectionist models is to make explicit the underlying differential equations, whereas they remained implicit in Thom's and Zeeman's works. These equations (introduced by Jack Cowan, Hugh Wilson and John Hopfield[39]) have very interesting properties and look very similar to those found in statistical physics in the theory of spin glasses. This has allowed, during the 1980s, a massive transfer of a large bulk of results from statistical physics to connectionist models.

But the fundamental limit of connectionist models is that they do not model the bifurcations of attractors that can result from an external dynamics modifying the attractors. They do use external dynamics but chiefly for modelling learning processes. The consequence is that they cannot afford models for constituent structures needed by all syntactic theories. A sharp debate took place at the end of the 1980s between classic cognitivism (Jerry Fodor and Zenon Pylyshyn) and connectionist cognitivism (Paul Smolensky). Fodor and Pylyshyn's thesis was that if one models the components of a sentence by attractors of a neural network, then it is not possible to model constituency. They were right. In order to model syntax, a model needs to be able to model constituency, which is impossible with attractors only.

However, as I wrote[40], Thom's actantial models provided an answer at the beginning of the 1970's, to this key issue of the late 1980's! Indeed, thanks to their built-in bifurcations, these models allow for what I call an "attractor syntax". If one models constituents (for instance, actants) with the attractors of some network, then it is not possible to model the relations between these constituents (for instance, actantial relations in a verbal node) through the attractors of the same network. One needs interactions between attractors, that is bifurcations. Attractors' bifurcations allow for the dynamical modelling of verbal nodes and constituent structures. It was the central idea of Thom's actantial graphs we have already discussed.

[39] H.R. Wilson and J.D. Cowan (1972). Excitatory and inhibitory interactions in localized populations of model neurons, *Biophys. J.*, 12 (1972) 1–24. J. J. Hopfield (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the USA*, 79, 8 (1982) 2554–2558.

[40] (1991) Why Connectionism is such a Good Thing. A Criticism of Fodor's and Pylyshyn's Criticism of Smolensky, *Philosophica*, 47, 1 (1991) 49-79. (1994) Attractor Syntax: Morphodynamics and Cognitive Grammars, *Continuity in Linguistic Semantics*, (C. Fuchs et B. Victorri eds), Amsterdam, John Benjamins, 1994, 167-187. (1995) Morphodynamics and Attractor Syntax. Dynamical and morphological models for constituency in visual perception and cognitive grammar, *Mind as Motion*, (R. Port and T. van Gelder eds.), Cambridge, MA, MIT Press, 1995, 227-281. Articles summarized in *Cognitive Morphodynamics*.

Q. Have these models proved seminal? What are the actual research results that are based on your work in cognitive morphodynamics?

We already talked about phonetics. In actantial syntax, the most important works are those by my friends Wolfgang Wildgen and Per Aage Brandt. In the teams of Aarhus and Copenhagen Peer Bundgaard[41], Svend Østergaard and Frederik Stjernfelt have used morphodynamic models. In Paris, David Piotrowski, a structuralist in the line of Hjelmslev has elaborated upon my propositions and plans to use neuroimaging (EEG). He claims that good neuroimaging experiments may help decide between linguistic theories since acceptability may be tested with neural waves, in particular N400[42].

Again in the field of linguistics, there are works by Bernard Victorri on synonymy that use dynamic models in an innovative way[43]. About prepositions, there are many works that still need to be modelled, in particular those by Claude Vandeloise[44]. In the volume edited in tribute to Vandeloise, there is a very interesting paper by Langacker[45].

In the field of perception, perceptive bifurcations have been studied extensively. There are models that follow Thom explicitly, others that follow Prigogine, and others synergetics; however, all these models are based on bifurcations. There is a large amount of empirical data. For instance, the Necker cube (figure 6), with its well-known double perspective. The same bi-dimensional stimulus can be interpreted as a tri-dimensional object in two different ways, and these two ways are bifurcating one in the other in a spontaneous and alternating manner along temporal series that have been studied in depth. The inversion of perspective is easy to understand. In bi-dimensional images, there are two points particularly salient and informative (the two edges in the centre of the figure); and according to the way you focus on one or the other of these two points, the cube is seen under one or the other perspective. There is also the example of the Rubin's face (figure 7) [46].

[41] Cf. for instance P. Bundgaard and J. Petitot (eds), (2010) *Aesthetic Cognition*, Special Issue of *Cognitive Semiotics*, 5, 2010. F. Stjernfelt and P. Bundgaard (eds) (2011) *Semiotics. Critical Concepts in Language Studies*, New York, Routledge.

[42] Piotrowski, D. (2009) *Phénoménalité* et *objectivité linguistiques*, Champion, Paris.

[43] See Victorri B., Fuchs C. (1996), *La polysémie. Construction dynamique du sens*. Paris, Hermès.

[44] Vandeloise, C. (1986) *L'Espace en Français: Sémantique des prépositions spatiales*, Paris, Editions du Seuil. (2009) "The genesis of spatial terms", *Language, Cognition and Space: the State of the Art and New Directions*, (V. Evans, P. Chilton eds), London, Equinox (*Advances in Cognitive Linguistics*), 157-178.

[45] Langacker, R. (2010) Reflections on the Functional Characterization of Spatial Prepositions, *Espace, Préposition, Cognition. Hommage à Claude Vandeloise*, (G. Col, C. Collin, eds), Corela.

[46] The vase-face by Edgar Rubin (Rubin, 1921) shows the importance of the figure–ground contrast in perception. According to whether one looks at the white area as the ground or the form, one sees two faces or a vase.

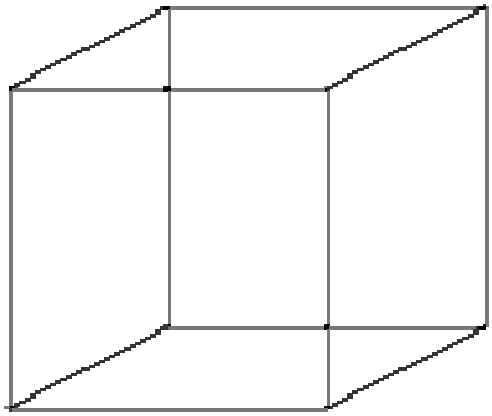


Figure 6. Necker's cube
(source: http://fr.wikipedia.org/wiki/Cube_de_Necker)

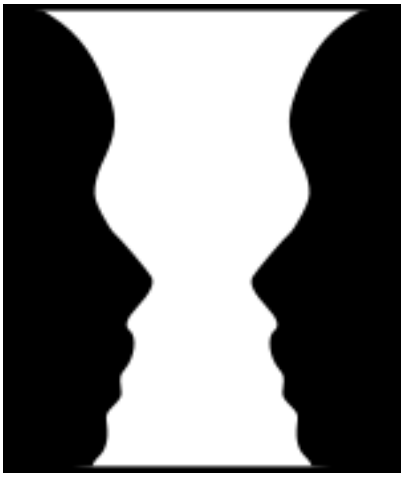


Figure 7. Rubin's face
(source: http://fr.wikipedia.org/wiki/Perception_figure-fond)

Thus, in many domains where mereological concepts of morphology and of structure mean something, one of the major issues is understanding how categories can emerge in continuous substrates. For this, one needs models where, in one way or another, there exist processes that produce discontinuities. One cannot escape this necessity and this explains the relevance of morphodynamical models.

Book review

Altmann, Gabriel, Köhler, Reinhard, *Forms and Degrees of Repetition in Texts. Detection and Analysis* (Quantitative Linguistics 68). Berlin/Munich/Boston: de Gruyter, 2015. ISBN 978-3-11-041179-9, viii+212 pp.
Reviewed by **Ján Mačutek**
Department of Applied Mathematics and Statistics, Comenius University, Mlynská dolina, SK - 84248 Bratislava, Slovakia, e-mail: jmacutek@yahoo.com

The book under review is, in fact, an updated, enlarged, and translated version of the older one by Altmann (1988), which was published in German. The updates ensure that it presents the current state of the art, and the translation makes it available to a much broader readership. As the title indicates, the book focuses on repetitions of textual units, where the word repetition is meant in a very general sense. The authors distinguish eight types of repetitions (see below), although it goes without saying that other possibilities remain open.

The book starts with an introduction, where methodology and the aims are briefly exposed. As is usual in quantitative linguistics (although still not in linguistics in general), definitions used do not have the ambition to capture the substance, they are operational (cf. Altmann 1997). Several reasons for repetition of linguistic units are discussed, e.g., a limitation of a unit inventory (which is valid especially for "lower" units, like phonemes and graphemes), grammar (a frequent use of function words), stylistic factors (refrain-like verses in some poems), etc. The book understands itself primarily as a step towards a (future) text theory; naturally, applications in different fields (authorship attribution, forensic linguistics, etc.) are not excluded.

The first – and at the same time the longest – chapter is dedicated to (as the authors name it) a "shapeless" repetition, or, in other words, to frequencies of multiple occurrences of units, with no other constraints laid on the units or their properties. This is the most common and the most simple approach to investigations of a repetition (especially word frequencies were investigated, cf. Baayen 2001 and Popescu et al. 2007, but also other units, like, e.g., graphemes, morphemes, etc.). Therefore, the book does not (and cannot) cover the entire spectrum of frequency studies, it limits itself to several topics from four areas: 1) comparisons of an observed unit frequency with its expected value, 2) comparing and testing differences among frequencies of several units, 3) global indicators of texts based on frequencies of all units which occur in a text (e.g., entropy or moments of a distribution), 4) regularities in frequency structures, a study of which can lead to a "discovery" of a text law (most often expressed by a probability distribution or by a function). All topics are accompanied by examples.

The second chapter focuses on positional repetitions, i.e., on unit frequencies at a particular position (which can be, e.g., at the beginning or at the end of a word, sentence, verse, etc.). The phenomenon is well known especially in poetic texts in form of rhyme, but also alliteration or assonance fall into this category. In fact, all examples in this chapter come from poems. First, several properties of rhyme are considered. Then word length climax in verses (i.e., the tendency of word length to increase from one position to the following one) is scrutinized and modelled by three different functions.

An associative repetition is defined as a (significant) coincidence of two units in a given frame, e.g., the coincidence of two words in a sentence. Only the coincidence of words (in verses and sentences) is presented, although there are other possibilities. Some methodological problems (e.g., should synonyms be considered as different words, or as instants of the same unit?) are listed, and several mathematical formalizations are suggested (probability distributions, graph theory).

Runs or iterative repetitions are uninterrupted sequences of identical units (or, in a more general form, of units of the same type). The most simple case is a binary sequence, which can be constructed for all linguistic units if a dichotomy is taken into account (units satisfying a particular condition versus the others). Statistical tests for determining whether runs are random or not and for comparing runs in two texts are presented. The authors mention briefly also more complex sequences consisting of more than two types of units.

An aggregative repetition is a generalization of the iterative repetition. Some units (e.g., words) almost never form uninterrupted sequences, but, on the other hand, they occur frequently in some places of a text and seldom in other places (i.e., they form clusters or aggregations). Similarly to the iterative repetitions, the problem is simpler if a binary sequence is considered. In such a case, the first question to be answered is whether distances between identical units are random or not; if not, the "mechanism" which generates them is described by Markov chains of different orders. Also here the models for sequences consisting of more than two types of units are more complicated.

The aggregative repetition, being a generalization of the iterative repetition, can itself be further generalized, if one considers not only appearances of identical units, but also of similar ones. This approach requires a definition of similarity. Phonetic similarity of verses is presented as an example.

The next chapter concentrates on repetitions in blocks. Unit frequencies display a regular behaviour if a text is segmented into blocks of equal length. The model for distributions in blocks is known as the Frumkina law (cf. Altmann & Burdinski 1982). Several mathematical models for the phenomenon are discussed.

A parallel repetition, which can be considered as a special case of the positional repetition, is the occurrence of equal (or similar) units in places which are parallel with respect to a "higher" frame, e.g., rhyme in a pair of verses. It is exemplified on texts from folk poetry.

Finally, a cyclic repetition, a wave-like structure, is common in poetry (especially sequences of stressed and unstressed syllables are often represented by oscillating curves), but also in other types of texts (an example of a sentence length sequence from a prosaic text is given). The Fourier analysis is mentioned as one of possibilities which can be used for the mathematical modelling of such repetitions.

There is an aspect in which the book is unsatisfying to say the least, namely, the references. Many works which are cited in the text cannot found in the list of cited literature (e.g., p. 1, Hřebíček 1997, Hřebíček 2000; p. 2, Searle 1969; p. 4, Altmann 2009, etc. – there are literally several dozens of such cases, at least 50 on the first 100 pages of the book). Furthermore, some other mistakes can be found the list of references (e.g., a wrong year).

In spite of this fact, the book can be recommended for researchers and students who work in the field of (quantitative) linguistics. It summarizes results known so far, and, although it is not a textbook, it provides a reasonable description of methods applied and a lot of examples. It can be useful for students of linguistics in quite a general sense, not only for students attending linguistic courses, but also for researchers coming from other educational backgrounds and now working in this interdisciplinary area. For an "established quantitative linguist" it gives an impetus towards a further development of a linguistic theory, as, in addition to achievements, it contains also a lot of challenges.

References

Altmann, G. (1988). *Wiederholungen in Texten*. Bochum: Brockmeyer.
Altmann, G. (1997). The nature of linguistic units. *Journal of Quantitative Linguistics, 3,* 1-7.

Altmann, G., & Burdinski, V. (1982). Towards a law of word repetitions in text-blocks. *Glottometrika, 4*, 146-167.

Baayen, R.H. (2001). *Word Frequency Distributions*. Dordrecht: Kluwer.

Popescu, I.-I., Grzybek, P., Jayaram, B.D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., & Vidya, M.N. (2009). *Word Frequency Studies*. Berlin, New York: de Gruyter.

Book Review

Zörnig, Peter; Stachowski, Kamil; Popescu, Ioan-Iovitz; Mosavi Miangah, Tayebeh; Chen, Ruina; Altmann, Gabriel, *Positional Occurrence in Texts: Weighted Consensus Strings.* Lüdenscheid: RAM-Verlag 2016. ISBN 978-3-942303-37-8, II + 179 pp.
Reviewed by **Emilia Bruch-Nemcova**

The book brings a new look at texts. Usually one analyzes texts sequentially or choosing the entities contained in it, forms sets and orders them in some way. One can study runs, distances, repetitions, etc. All these aspects may be quantified and measured, models can be proposed and subsumed in a background theory. The majority of articles concerning texts are either global or linear views.

The authors of the above book show that text can be considered also vertically. But if so, it must be subdivided in some similar, "natural" or defined entities. Since written texts have punctuation, they may be subdivided in sentences. In the sentence one takes into account the position and computes the number of the given entity in that position. The position may display a tendency which is expressed by the maximum proportion of the column. The authors consider positionally the following entities: word length in syllables, polysemy, frequencies, parts of speech, canonical syllable types (CV, CVC,…) in several texts in seven languages: Chinese, French, German, Persian, Polish, Slovak and Turkish. Since all numbers and fitted models are presented, the book is full of tables displaying the empirical and theoretical distributions. Each data are presented also graphically. Needless to say, the models fit well but if further fifty languages will be scrutinized, one must strive for a very general theory. Up to now, they use 12 models, all of which can be found in the *unified theory*. The differences are evidently caused by boundary conditions which must be scrutinized individually and their source must be found in each individual case separately. It may be a sign of style or age or development, etc.

The authors proceed in a well defined "quantitative" way: first a hypothesis is presented, then it is derived from theoretical conjectures by means of difference equations and at last, the hypothesis is tested. The book shows that even the parameters of the acquired models are linked with one another – good sign of self-regulation.

The texts are further characterized – on the basis of Concensus Strings – by Ord's criterion, non-smoothness indicator, von Neumann's mean square successive difference and the roughness indicator, all of which can be found in quantitative textological works. All basic data are presented in the *Appendix*, hence the reader can use them both for checking the results and for further investigations. If a new language is added and its text analyses strongly differ from those already obtained, one will be forced to re-think the theoretical background, as is usual in all sciences.

The authors cared for clear explication and simplicity of statistical procedures hence even a beginner can apply everything to other texts, text types or languages. There are two approaches: static and dynamic, both showing different views.

The vertical structuring of the text may be studied also applying other basic units, e.g. clauses, verses, strophes, Frumkina's sections, etc. An application to poetry would surely bring new vistas and one could find links to the rhythmical background (if it exists) or show that this aspect is fully independent. The most important problem is the a priori stating of hypotheses, measurement and testing. The last step, construction of a theory is always a task for the future.

Book Review

Radek Čech: *Tematická koncentrace textu v čestine (Thematic concentration of the text in Czech).* Praha: Ústav fomální a aplikované lingvistiky (= Studies in Computational and Theoretical Linguistics) 2016, 236 pp.
Reviewed by **Hanna Gnatchuk**

This important book is, unfortunately, written in Czech, hence not accessible to all quantitative linguists. Nevertheless, it discusses one of the modern concepts, introduced and analyzed in the last years. The author is aware of all problems associated with the introduction (definition, formalization, measurement,…) of any concept into linguistics and performs all necessary steps, criticizes the concept, shows its testability and various other possibilities. The book shows clearly that mathematical models must be considered only as a step in the evolution of science. They are never the final truth, there are always different ways to attain a result. After having defined the concept of thematic concentration and the possibility of its testing, the author presents other analogous methods, so to say, variations. He abandoned the methodological error of considering a unit as something "given" – a phenomenon still surviving in several sciences – and compares various alternatives. The methodological background is discussed in detail. This is a very positive signal: one begins to take into account the philosophy of science.

A chapter is devoted to the difficult problem of removing the influence of text size but even here various possibilities are shown. The text can be analyzed also cumulatively or making windows, it is possible to consider perhaps the hrebs introduced by Hřebíček and applied many times, or Belza-chains analyzing the text sequentially. All this is discussed in the book.

As any linguistic entity, thematic concentration is no isolated property; it is linked with several other properties. Some of them belong to the Köhlerian control circuit which increases every year. The book scrutinizes the text richness expressed in different ways; the key-words in text and the associative thematic structure of the text. This all represents a very high conceptual abstraction level.

Thematic concentration can be utilized also for classifying texts, forming of text types, capturing the individualities of the style of an author. A special chapter is dedicated to very long texts. An interesting problem would be the comparison of texts representing the same text type in different languages; one could also consider the expression of this property in original and translated texts, especially poetic ones. The problem can be strongly extended in the future.

All this can be computed by means of the QUITA-software prepared by the author himself and is freely downloadable from the Internet
(https://code.google.com/archive/p/oltk/).

More than 60 pages are devoted to tables with numbers. The numbers have been won from 1168 Czech texts, an enormous number, if compared with other investigations.

All chapters contain the relevant formulas, tables presenting the numbers, and graphs accompanying each result. The showing of all numbers in a quantitative investigation is extremely important because this is the only way to check the results. The references contain both old works and the newest ones. The book is a basic source for entering this domain of research.

Bibliography of Word Length

Abbe, S. (2000). Word length distribution in Arabic letters. *Journal of Quantitative Linguistics 7, 121-127.*

Ahlers, A. (2001). The distribution of word length in different types of Low German texts. In: Best, K.-H. (ed.), *Häufigkeitsverteilungen in Texten: 43-58.* Göttingen:Peust & Gutschmidt.

Altmann, G. (2013). Aspects of word length. In: Köhler, R., Altmann, G. (eds.), *Issues in Quantitative Linguistics Vol 3: 23-38*. Lüdenscheid: RAM-Verlag.

Altmann, G., Best, K.-H. (1996). Zur Länge der Wörter in deutschen Texten. In: Schmidt, P. (ed.), *Glottometrika 15, 166-180.* Trier: WVT.

Altmann, G., Best, K.-H., Wimmer, G. (1997). Wortlänge in romanische Sprachen. In: Gather, A., Werner, H. (eds.), *Semiotische Prozesse und natürliche Sprache: 1-13.* Stuttgart: Steiner.

Altmann, G., Erat, E., Hřebíček, L. (1996). Word length distribution in Turkish texts. In: Schmidt, P. (ed.), *Glottometrika 15, 195-204*. Trier: WVT.

Ammermann, S. (1997). Untersuchung zur Wortlängehäufigkeit in Briefen Kurt Tucholskys. In: Best, K.-H. (ed.), *Glottometrika 16, 63-70*. Trier: WVT.

Ammermann, S. (2001). Zur Wortlängenverteilung in deutschen Briefen über einen Zeitraum von 500 Jahren. Best, K.-H. (ed.), *Häufigkeitsverteilungen in Texten: 59-91*. Göttingen:Peust & Gutschmidt.

Ammermann, S., Bengtson, M. (1997). Zur Wortlängenhäufigkeit im Schwedischen: Gunnar Ekelöfs Briefe. In: Best, K.-H. (ed.), *Glottometrika 16, 88-97*. Trier: WVT.

Antić, G., Kelih, E., Grzybek, P. (2006). Zero-syllable words in determining word length. In: Grzybek, P. (ed.) (2006)*: 117-156.*

Arlt, I. (2006). Zur Wortlängenverteilung in SMS-Texten. *Göttinger Beiträge zur Spraxchwissenschaft 13, 9-21*.

Balschun, C. (1997). Wortlängenhäufigkeiten in althebräischen Texten. In: Best, K.-H. (ed.), *Glottometrika 16, 174-179*. Trier: WVT.

Barbaro, S. (2000). Word length distribution in Italian letters by Pier Paolo Pasolini. *Journal of Quantitative Linguistics 7, 115-120.*

Bartels, O., Strehlow, M. (1997). Zur Häufigkeit von Wortlängen in deutschen Briefen im 19. Jahrhundert und in der ersten Hälfte des 20. Jahrhunderts (Bismarck, Brecht, Kafta, T. Mann, Tucholsky). In: Best, K.-H. (ed.), *Glottometrika 16, 71-76*. Trier: WVT.

Bartens, H.-H., Best, K.-H. (1997). Word length distribution in Sámi texts. *Journal of Quantitative Linguistics 4. 45-52.*

Bartens, H.-H., Best, K.-H. (1996). Wortlängen in estnischen Texten. *Ural-Altaische Jahrbücher N.F.- 14, 112-128.*

Bartens, H.-H., Best, K.-H. (1997). Wortlängen in erzamordwinischen Texten. *Linguistica Uralica 23, 5-13.*

Bartens, H.-H., Best, K.-H. (1997). Wortlängen in Tscheremissischen (Mari). *Finnish-Ugrische Mitteilungen 20, 1-20.*

Bartens, H.-H., Zöbelin, Th. (1997). Wortlängenhäufigkeiten im Ungarischen. In: Best, K.-H. (ed.), *Glottometrika 16, 195-203*. Trier: WVT.

Bartkowiakowa, A., Gleichgewicht, B. (1962). O dlugości sylabicznej wyrazów w tekstach autorów polskich. *Zastosowania matematyki 6, 309-319.*

Bartkowiakowa, A., Gleichgewicht, B. (1964). Zastosowanie, dwuchparametrowych rozkladów Fuchsa do opisu dlugości sylabicznej warazów w różnych utworach autorów polskich. *Zastosowania matematyki 7, 345-352.*

Bartkowiakowa, A., Gleichgewicht, B. (1965). O rozladoch dlugości sylabicznej wyrazów v różnych tekstach. In: Mayenowa, M.R. (ed.), *Poetyka i matematyka: 164-173.* Warszawa.

Becker, C. (1996). Word lengths in letters oft he Chilenic author Gabriela Mistral. *Journal of Quantitative Linguistics 3, 128-131.*

Behrmann, G. (1997). Die Wortlängenhäufigkeit von deutschsprachigen naturwissenschaftlichen Publikationen**.** In: Best, K.-H. (ed.), *Glottometrika 16, 77-87*. Trier: WVT.

Best, K.-H. (1996). Word length in Old Icelandic songs and prose texts. *Journal of Quantitative Linguistics 3, 97-105.*

Best, K.-H. (1996). Zur Bedeutung von Wortlängen, am Beispiel althochdeutscher Texte. *Papiere zur Linguistik 55, 141-152.*

Best, K.-H. (1996). Zur Wortlängenhäufigkeit in schwedischen Pressetexten. In: Schmidt, P. (ed.), *Glottometrika 15, 147-157*. Trier: WVT.

Best, K.H. (1997). Warum nur Wortlänge? Nicht nur ein Vorwort. In: Best, K.-H. (ed.), *Glottometrika 16, V-XII.* Trier: WVT.

Best, K.-H. (1997). Zur Wortlängenhäufigkeit in deutschsprachigen Pressetexten. In: Best, K.-H. (ed.), *Glottometrika 16, 1-15.* Trier: WVT.

Best, K.-H. (1997). Wortlängen in mittelhochdeutschen Texten. In: Best, K.-H. (ed.), *Glottometrika 16, 40-54.* Trier: WVT.

Best, K.-H. (2000). Wie viele Morphe enthalten deutsche Wörter? Am Beispiel einiger Fabeln Pestalozzis. In: Ondrejovič, S., Považajová, M. (eds.), *Lexicographica `99:* Bratislava: Veda.

Best, K.-H. (2001). Wortlängen in Texten gesprochener Sprache. *Göttinger Beiträge zur Sprachwissenschaft 6.*

Best, K.-H. (2005). Wortlängen. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistikcs. An International Handbook: 260-273.* Berlin: de Gruyter.

Best, K.-H. (2010). Silben-, Wort- und Morphlängen bei Lichtenerg. *Glottometrics 21, 1-13.*

Best, K.-H., Altmann, G. (1996). Zur Länge der Wörter in deutschen Texten. In: Schmidt, P. (ed.), *Glottometrika 15: 166-180*. Trier: WVT.

Best, K.-H., Brynjólfsson, E. (1997). Wortlängen in isländischen Briefen und Pressetexten. *Skandinavistik 27, 24-40.*

Best, K.-H., Kaspar, I. (1998). Wortlängen in färöischen Briefen. *Naukovyj Visnik Černivec`koho Universytetu 41. Hermans' ka filolohija 3-14.*

Best, K.-H., Kaspar, I. (2001). Wortlängen im Färöischen. *In:* Best, K.-H. (ed.), *Häufigkeitsverteilungen in Texten: 92-100.* Göttingen: Peust & Gutschmidt.

Best, K.-H., Medrano, P. (1997). Wortlängen in Ketschua-Texten. In: Best, K.-H. (ed.), *Glottometrika 16, 204-212*. Trier: WVT.

Best, K.-H., Özmen, E. (1996). Wortlängenhäufigkeiten in türkischen Texten und ihre linguistischen Implikationen. *Archiv orientální 64, 19-30.*

Best, K.-H., Song, H-Y. (1996). Wortlängen im Koreanischen. *Asian and African Studies 5, 39-49.*

Best, K.-H., Zhu, J. (1994). Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa (mit einem Ausblick auf das Chinesische). In: Klenk, U. (ed.), *Computatio linguae II: 19-30.* Stuttgart: Steiner.

Best, K.-H., Zhu, J. (2001). Wortlängen in chinesischen Texten und Wörterbüchern. In: Best, K.-H. (ed.), *Häufigkeitsverteilungen in Texten: 101-114.*. Göttingen: Peust & Gutschmidt.

Best, K.-H., Zinenko, S. (1998). Wortlängenverteilungen in Briefen A.T. Twardowskis. *Göttinger Beiträge zur Sprachwissenschaft 1, 7-19.*

Best, K.-H., Zinenko, S. (1998). Wortlängenverteilungen in Gedichten des ukrainischen Autors Ivan Franko. In: Genzor, J., Ondrejovič, S. (eds.), *Pange Lingua: 201-214*. Bratislava: Veda.

Best, K.-H., Zinenko, S. (1998). Wortkomplexität im Ukrainischen und ihre linguistische Bedeutung. *Zeitschrift für Slavische Philologie 58, 107-123.*

Best, K.-H., Zinenko, S. (2001). Wortlängen in Gedichten A.T.Twardowskis. In: Uhlířová, L., Wimmer, G., Altmann, G., Köhler, R. (eds.), *Text as a linguistic paradigm: Levels, constituents, constructs: 21-28.* Trier: WVT.

Brainerd, B. (1971). On the distribution of syllables per word. *Mathematical Linguistics 57, 1-18.*

Breiter, M.A. (1994). Length of Chinese words in relation to their other systemic features. *Journal of Quantitative Linguistics 1, 224-231.*

Bürmann, G., Frank, H., Lorenz, L. (1993). Informationstheoretische Untersuchungen über Rang und Länge deutscher Wörter. *Grundlagenstudien aus Kybernetik und Gesiteswissenschaften 4, 73-90.*

Cercvadze, G.N., Čikoidze, G.B., Gačećiladze, T.G. (1959). Primenenie matematičeskoj teorii slovoobrazovanija k gruzinskomu jazyku. *Soobščenija akademii nauk Gruzinskoj SSR 22(6), 705-710.*

Christiansen, B. (1997). Wortlängenverteilung in deutschsprachigen Barockgedichten. In: Best, K.-H. (ed.), *Glottometrika 16, 16-39.* Trier: WVT.

Dieckmann, S., Judt, B. (1996). Untersuchung zur Wortlängenverteilung in französischen Pressetexten und Erzählungen. In: Schmidt, P. (ed.), *Glottometrika 15, 158-165*. Trier: WVT.

Dittrich, H. (1996). Word length frequency in the letters of G.E. Lessing. *Journal of Quantitative Linguistics 3, 260-264.*

Drechsler, J. (2001). Häufigkeitsverteilung von Wortlängen in gällischen Texten. In: Best, K.-H. (ed.), *Häufigkeitsverteilungen in Texten: 43-58.* Göttingen: Peust & Gutschmidt.115-123.

Egbers, J., Groen, C., Rauhaus, E., Podehl, R. (1997). Zur Wortlängenhäufigkeit in griechischen Koine-Texten. In: Best, K.-H. (ed.), *Glottometrika 16, 108-120*. Trier: WVT.

Elderton, W.P. (1940). A few statistics on the length of English words. *Journal of the Royal Statistical Society, series A 112, 436-445.*

Feldt, S., Janssen, M., Kuleisa, S. (1997). Untersuchung zur Gesetzmäßigkeit von Wortlängenhäufigkeiten in französischen Briefen und Pressetexten. In: Best, K.-H. (ed.), *Glottometrika 16, 145-151*. Trier: WVT.

Fickermann, I., Markner-Jäger, B., Rothe, U. (1984). Wortlänge und Bedeutungskomplexität. In: Boy, J., Köhler, R. (eds.), *Glottometrika 6: 115-126*. Bochum: Brockmeyer.

Fónagy, I. (1960). A szavak hossa a magyar beszédben. *Magyar Nyelvör 1960, 355-360.*

Frischen, J. (1996). Word ength analysis of Jane Austen's letters. *Journal of Quantitative Linguistics 3, 80-84.*

Fucks, W. (1955). *Mathematische Analyse von Sprachelementen, Sprachstil und Sprachen.* Köln, Oplades: Westdeutschwer Verlag.

Gaeta, L. (1994). Wortlängenverteilung in italienischen Texten. *Zeitschift für empirische Textforschung 1, 44-48.*

Girzig, P. (1997). Untersuchung zur Häufigkeit von Wortlängen in russischen Texten. In: Best, K.-H. (ed.), *Glottometrika 16, 152-162.*. Trier: WVT.

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

Grotjahn, R., Altmann, G. (1991). Modelling the distribution of word length: Some methodological problems. In: Köhler, R., Rieger, B. (eds.), *Contributions to Quantitative Linguistics: 141-153*. Dordrecht/Boston/London: Kluwer.

Grzybek, P. (1998). Explorative Untersuchung zur Wort- und Satzlänge kroatischer Sprichwörter (Am Beispiel der `Poslovice` von Duro Daničić, 1971). In: Nikolaeva, T.M. (ed.), *Polytropon: 447-465*. Moskva: Indrik.

Grzybek, P. (2000). Pogostnostna analiza besed iz elektronskego korpusa slovenskih besedil. *Slavistična Revija Apf.-jun 2000, 141-157.*

Grzybek, P. (2006). History and methodology of word length studies. In: Grzybek, P. (ed.) *2006: 15-90.*

Grzybek, P. (ed.) (2006). *Contributions to the science of text and language. Word length studies and related issues*. Dordrecht: Springer.

Grzybek, P. (2011). Der Satz und seine Beziehungen. I: Satzlänge und Wortlänge im Russischen (Am Besipiel ovn L.N. Tolstojs „Anna Karenina"). *Anzeiger für Slawische Philologie 39. 39-74.*

Grzybek, P., Kelih, E., Stadlober, E. (2008). The relation between word length and sentence length. An intra-systemic perspective in the core data structure. *Glottometrics 16, 111-121.*

Hasse, A., Weinbrenner, M. (1997). Zur Häufigkeit von Wortlängen in englischen Texten. In: Best, K.-H. (ed.), *Glottometrika 16, 98-107*. Trier: WVT.

Hein, M. (1997). Wortlängen in Briefen des spanischen Dichters Federico García Lorca. In: Best, K.-H. (ed.), *Glottometrika 16, 138-144*. Trier: WVT.

Hollberg, C. (1997). Wortlängenhäufigkeiten in italienischen Pressetexten. In: Best, K.-H. (ed.), *Glottometrika 16, 127-137*. Trier: WVT.

Janssen, E., Suhren, S. (2000). Wortlängenhäufigkeiten in ostrfriesisch-niederdeutschen Gedichten von Hans-Hermann Briese. *Göttinger Beiträge zur Sprachwissenschaft 4, 53-62.*

Kahl, S. (2002). Wortlängenverteilungen in wogulischen Texten. *Göttinger Beiträge zur Sprachwissenschaft 7, 51-63.*

Kaydanova, L.(2004/5). Zur Wortlängenhäufigkeit in usbekischen Texten. *Göttinger Beiträge zur Sprachwissenschaft 10/11, 37-66.*

Kiefer, A. (2001). Wortlängenverteilung im Pfälzischen. In: Best, K.-H. (ed.), *Häufigkeitsverteilungen in Texten: 124-131.* Göttingen:Peut & Gutschmidt.

Kim, I., Altmann, G. (1996). Zur Wortlänge in koreanischen Texten. In Schmidt, P. (ed.), *Glottometrika 15: 205-213.* Trier: WVT.

Kiyko, S. (2007). Wortlängen im Gotischen. *Glottometrics 13, 47-58.*

Kiyko, S. (2007). Wortlängen im Weißrussischen. *Glottometrics 14, 46-57.*

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

Kuhr, S. (2001). Zur Wortlängenhäufigkeit in Luthers Liedern und Fabeln. In: Best, K.-H. (ed.), *Häufigkeitsverteilungen in Texten: 133-141*. Göttingen: Peust & Gutschmidt.

Kuhr, S., Müller, B. (1997). Zur Wortlängenhäufigkeit in Luthers Briefen. In: Best, K.-H. (ed.), *Glottometrika 16, 55-62*. Trier: WVT.

Laass, F. (1996). Zur Verteilung von Wortlängen in deutschen Lesebuchtexten. In: Schmidt, P. (ed.), *Glottometrika 15, 181-194*. Trier: WVT.

Martin, W. (1976). On the evolution of word-length in Dutch. In: Jones, A., Churchhouse, R.F. (1976), *The Computer and Literary and Linguistic Studies: 271-284.* Cardiff: The University of Cardiff Press.

Marx, M. (2001). Zu den Wortlängen in polnischen Briefen. *Glottometrics 1, 52-62*.

Meyer, P. (1997). Word length distribution in Inuktitut narratives: Empirical and theoretical findings. *Journal of Quantitative Linguistics 4, 143-155.*

Meyer, P. (1999). Relating word length in morphemic structure: A morphologically motivated class of probability distributions. *Journal of Quantitative Linguistics 6, 66-69.*

Müller, F. (2003). Wortlängen in finnischen E-Mails und Briefen. *Göttinger Beiträge zur Sprachwissenschaft 8, 71-85.*

Nemcová, E., Atmann, G. (1994). Zur Wortlänge in slowakischen Texten. *Zeitschrift für empirirische Textforschung 1, 40-43.*

Popescu, I.-I., Best, K.-H., Altmann, G. (2014). *Unified modeling of length in language*. Lüdenscheid: RAM-Verlag.

Rheinländer, N. (2001). Die Wortlängenhäufigkeit im Niederländischen. In: Best, K.-H. (ed.), *Häufigkeitsverteilungen in Texten: 142-152.* Göttingen: Peust & Gutschmidt.

Riedemann, G. (1997). Wortlängenhäufigkeiten in japanischen Pressetexten. In: Best, K.-H. (ed.), *Glottometrika 16, 180-184*. Trier: WVT.

Riedemann, H. (1996). Word length distribution in English press texts. *Journal of Quantitative Linguistics 3, 265-271*.

Röttger, W. (1996). The distribution of word length in Ciceronian letters. *Journal of Quantitative Linguistics 3, 68-72.*

Röttger, W., Schweers, A. (1997). Wortlängenhäufigkeit in Plinius-Briefen. In: Best, K.-H. (ed.), *Glottometrika 16, 121-126*. Trier: WVT.

Rottmann, O. (1997). Word-length counting in Old Church Slavonic. *Journal of Quantitative Linguistics 4, 252-256.*

Rottmann, O.A. (1999). Word and syllable length in East Slavonic. *Journal of Quantitative Linguistics 6, 235-238.*

Rottmann, O.A. (2003). Word lengths in the Baltic languages – are they of the same type as the word lengths in Slavic languages? *Glottometrics 6, 52-60*

Stark, A.B. (2001). Die Verteilung von Wortlängen in schweizerdeutschen Privatbriefen. In: Best, K.-H. (ed.), *Häufigkeitsverteilungen in Texten: 153-161*. Göttingen:Peut & Gutschmidt.

Uhlířová, L. (1995). O jednom modelu rozložení délky slov. *Slovo a slovesnost 56. 8-14.*

Uhlířová, L. (1996). On the generality of statistical laws and individuality of texts. A case of syllables, word forms, their length and frequencies. *Journal of Quantitative Linguistics 2, 238-247.*

Uhlířová, L. (1996). How long are words in Czech?. In: Schmidt, P. (ed.), *Glottometrika 15: 134-146. Trier: WVT.*

Uhlířová, L. (1997). Word length distribution in Czech: On the generality of linguistic laws and individuality of texts. In: Best, K.-H. (ed.), *Glottometrika 16, 163-173*. Trier: WVT.

Uhlířová, L. (1999). Word length modeling: Intertextuality as a relevant factor? *Journal of Quantitative Linguistics 6, 252-256.*

Uhlířová, L. (2001). On word length, clause length and sentence length in Bulgarian. In: Uhlířová, L., Wimmer, G., Altmann, G., Köhler, R. (eds.), *Text as a linguistic paradigm: Levels, Constituents, Construct: 266-282.* Trier: WVT.

Vettermann, A., Best, K.-H. (1997). Wortlängen in Finnischen. *Suomalais-ugrilaisen seuran aikakauskirja/Jounral de la Societé Finno-Ougrienne 87, 249-262*.

Wilson, A. (2003). Word-length distribution in modern Welsh prose. *Glottometrics 6, 35-39.*

Wilson, A. (2006). Word-length distribution in present-day Lower-Sorbian newspaper letters. In: Grzybek, P. (ed.) *2006: 319-327.*

Wilson, A., McEnery, T. (1998). Word length distribution in Biblical and Medieval Latin. *The Prague Bulletin of Mathematical Linguistics 70, 5-21.*

Wimmer, G., Altmann, G. (1996). The theory of word length: Some results and generalizations. In: Schmidt, P. (ed.), *Glottometrika 15, 112-133.* Trier: WVT.

Wimmer, G., Köhler, R., Grotjahn, R., Altmann, G. (1994). Towards a theory of word length distribution. *Journal of Quantitative Linguistics 1, 98-106.*

Zhu, J., Best, K.-H. (1992). Zum Monosyllabismus im Chinesischen. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 45, 341-355.*

Zhu, J., Best, K.-H. (1992). Zum Wort I modernen Chinesischen. *Oriens Extremus 35, 45-60.*

Zhu, J., Best, K.-H. (1997). Zur Modellierung der Wortlängen im Chinesischen. In: Best, K.-H. (ed.), *Glottometrika 16, 185-194*. Trier: WVT.

Zhu, J., Best, K.-H. (1997). Wortlänge in chinesischen Briefen. In: Hongjun Cai (ed.), *Neue Forschungen chinesischer Germanisten in Deutschland: 121-129.* Frankfurt: Peter Lang.

Zhu, J., Best, K.-H. (1998). Wortlängenhäufigkeiten in chinesischen Kurzgeschichten. *Asian and African Studies 7, 45-51.*

Ziegler, A. (1996). Word engths distribution in Brazilian-Portuguese texts. *Journal of Quantitative Linguistics 3, 73-79.*

Ziegler, A. (1998). Word length in Portuguese texts. *Journal of Quantitative Linguistics 5, 115-120.*

Zuse, M. (1996). The distribution of word length in English letters of Sir Philip Sidney. *Journal of Quantitative Linguistics 3, 272-276.*

Other linguistic publications of RAM-Verlag:

Studies in Quantitative Linguistics

Up to now, the following volumes appeared:

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, VI + 148 pp.
15. K.-H. Best, E. Kelih (Hrsg.), *Entlehnungen und Fremdwörter: Quantitative Aspekte.* 2014, IV + 163 pp.
16. I.-I. Popescu, K.-H. Best, G. Altmann, *Unified modeling of length in language.* 2014. III + 123 pp.
17. G. Altmann, R. Čech, J. Mačutek, L. Uhlířová (eds.), *Empirical approaches to text and language analysis.* 2014, IV + 230 pp.
18. M. Kubát, V. Matlach, R. Čech, *QUITA. Quantitative Index Text Analyzer*. 2014, IV + 106 pp.
19. K.-H. Best (Hrsg.), *Studies zur Geschichte der Quantitativen Linguistik. Band 1.* 2015, III + 159 pp.
20. P. Zörnig et al., *Descriptiveness, activity and nominality in formalized text sequences.* IV+120 pp.
21. G. Altmann, *Problems in Quantitative Linguistics Vol. 5.* 2015, III+146 pp.
22. P. Zörnig et al. (2016). *Positional occurrences in texts: Weighted Consensus Strings.* 2016. II+179.pp