Glottometrics 14, 2007, 58-71

On the dynamics of word classes in text

Ioan-Iovitz Popescu, Bucharest[1]
Karl-Heinz Best, Göttingen
Gabriel Altmann, Lüdenscheid

Abstract: In this study, the distributions of certain parts of speech, especially auxiliaries, is investigated. Using the definition of the *h*-point, we define the thematic concentration of the text and introduce the concentration unit *tcu*.

Keywords: Word classes, auxiliaries, thematic concentration, distributions

1. Introduction

The dynamics of word classes in text can be evaluated in different ways. Firstly, according to their frequency distribution, which follows a special probability distribution (cf. Hammerl 1990; Best 1994, 1997, 2000, 2001; Schweers, Zhu 1991; Ziegler 1998, 2001); secondly, according to sequences of individual classes in text building a special distance pattern or a time series (cf. Pawlowski 2005; Ziegler, Best, Altmann 2001, 2002) and thirdly, according to the participation of a word class in a frequency class, an aspect which can be characteristic both of texts and languages. This last mode of evaluation will be scrutinized in this paper.

Consider the frequency distribution of words in a text, ranked according to decreasing frequency. For the sake of simplicity we shall differentiate only auxiliaries (function words, synsemantics) and autosemantics (nouns, verbs, adjectives, adverbs, numerals). This differentiation is not crisp and there are different classifications even in one language because language does not care for classes; they are no more than our conceptual creations, mostly based on Latin. If we look at words occurring exactly once in the text (g(1)), we state that the proportion of auxiliaries in this frequency class is very small. Taking class g(2) of words occurring twice, the proportion of auxiliaries increases. Continuing to the most frequent class we can observe that the proportion of auxiliaries grows monotonously and in some class it attains 1.00. That means that in long texts the curve of proportions of auxiliaries begins with zero (or a number very near zero) and in the class of the most frequent word – which is usually an auxiliary – it must be 1.00. Now, since the curve increases slowly at the beginning and approaches 1.00 long before it attains it, its form must change from convex to concave, i.e. at least for auxiliaries it must have an inflection point. This assumption can easily be translated in mathematics. Let y_x be the proportion of auxiliaries in the frequency class x. Then the rate of change of y_x is proportional to its own height and to its distance to 1, i.e.

$$\frac{dy}{dx} = ay(1-y)\,, \tag{1}$$

whose solution yields

[1] Address correspondence to: iovitzu@gmail.com

(2) $$y = \frac{1}{1 + be^{-ax}}$$

known as the logistic curve and from historical linguistics as Piotrowski law. Up to now the curve has been used mostly in historical linguistics where a class of changes has this form (there are also incomplete and reversible changes); its appearance in text structuring is a surprise. In the present article we shall analyze only the two fuzzy classes of auxiliaries and autosemantics. Many words vary in their class membership. Consider the German morpheme "auf". It can be a preposition ("auf dem Haus" – on the house), a fixed prefix ("Auftrag" – assignment), a prefix followed by another prefix ("aufgewendet" – spent), or a verbal particle with the status of an adverb ("stehe auf!" – stand up!), although not all grammarians would agree with this last. If one uses a PoS-tagger, the program represents the grammatical philosophy of the programmer/linguist. It is not our aim to emphasize a particular grammar but to show that under the given conditions "grammar words" have a certain dynamic.

2. Application

Consider first the distribution of word forms in A. v. Droste-Hülshoff´s "Der Geiergriff" as shown in the first two columns of Table 1. Here $g(x)$ is the number of words occurring exactly x-times; *Aux* is the number of auxiliaries in $g(x)$, $p(Aux)$ is the proportion of auxiliaries and $p(Aux)_{theor}$ is the computed value from the fitting logistic curve, Eq.2.

.

Table 1
Frequency distribution of word forms in Droste-Hülshoff

x	***g(x)***	***Aux***	***p(Aux)***	***p(Aux)$_{theor}$***
1	380	23	0.061	0.114
2	71	7	0.099	0.178
3	14	4	0.286	0.265
4	11	4	0.364	0.376
5	6	4	0.667	0.501
6	8	5	0.625	0.626
7	4	3	0.75	0.737
8	2	4	0.5	0.824
9	1	9	1	0.886
10	1	10	1	0.929
11	2	11	1	0.956
12	2	12	1	0.973
16	2	16	1	0.996
17	1	17	1	0.998
20	1	20	1	1
26	1	26	1	1
36	1	36	1	1
39	1	39	1	1

In Figure 1 a trend can be clearly seen. The formula obtained by optimization is

$p_{theor} = 1/(1 + 12.9149 \exp(-0.5125\, x))$

and the determination coefficient is $R^2 = 0.92$. Hence we can conclude that at least in this text the auxiliaries have the frequency structure (2). The inflection point of (2) is $x = (\ln b)/a$, i.e. in the above case $x = (\ln 12.9149)/ 0.5125 = 4.99 \approx 5$.

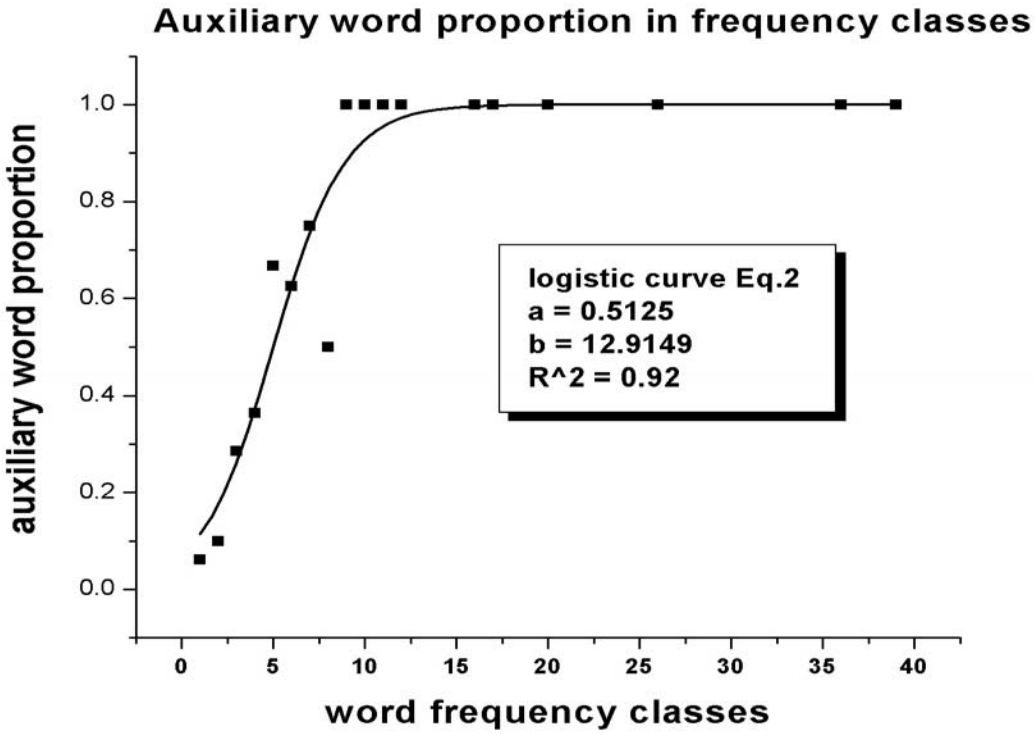

Figure 1. The frequency structure of auxiliaries in A. Droste-Hülshoff´s text

In the same way, we analysed some other short German texts and obtained the results in Table 2. In some cases, when the subject of the text is named very frequently (e.g. "Becher" in Schiller's "Taucher"), the given noun gets under the auxiliaries and if there are no other words of the same frequency, we obtain an observed $y = 0$ in that point. Such points are left out because they were "reserved" for non-auxiliaries. This happens in rather shorter texts. The curve has in all cases the same course and displays an inflection point (*IP*) which can be considered a characteristic of the text. Since the inflection point is a function of the two parameters ($x = (\ln b)/a$), it is perhaps sufficient to use them to characterize the text. In the texts of Table 2, the *IP* points are relatively stable.

Table 2
Frequency structure of auxiliaries in some short German texts

Author	Text	*b*	*a*	R^2	*IP*
Goethe	Elegie 19	5.0590	0.5776	0.63	2.81
Anonym	Mäuschen	5.0833	0.3261	0.85	4.99
Goethe	Der Gott und die Bajadere	10.0648	0.7103	0.94	3.25
Droste-Hülshoff	Der Geiergriff	12.9121	0.5125	0.92	4.99
Anonym	Zaunbär	18.4856	0.5314	0.86	5.49
Möricke	Peregrina	19.2843	0.8214	0.95	3.60
Schiller	Der Taucher	34.2635	0.8656	0.95	4.08

The same procedure can be performed not only for formal classes, whether isolated or pooled, but also for semantic classes like "processual expressions" containing verbs expressing some activity (no verbs like "sleep", "be", "have" etc.) and also nouns derived from activity verbs or even adjectives (e.g. "donnernder Huf" - thundering hoof). In English a number of conversions belong to this class. According to our hypothesis, the frequency distribution of words (in this case, rather, word forms) is semantically structured; and even individual frequency classes have their particular semantic spectrum, changing from class to class.

Let us now consider an English text, namely Rutherford's Nobel lecture. The PoS-tagging of the text has been performed with the aid of the CLAWS tagger (http://www.comp.lancs.ac.uk/ucrel/trial.html) and the frequency count with the aid of a very reliable counter that can be found at http://www.georgetown.edu/faculty/ballc/webtools/web_freqs.html. Some small uncertainties of the tagger have been corrected by hand. The following "parts of speech" have been pooled in one class: articles (ATO), adverbs (AVO, AVP, AVQ), conjunctions (CJC, CJS, CJT), determiners (DPS, DTO, DTQ), existential "there" (EXO), pronouns (PNI, PNP, PNQ, PNX, POS), prepositions (PRF, PRP), the infinitive marker "to" (TOO), the negation (XXO). In the same way as above, the participation of this class in frequency classes has been ascertained and we obtained the result given in Table 3. The classes x = 1,..,20 have been left separated, classes 21-100 have been pooled and the mean x = 60.5 has been taken; for the pooled classes 101-464 the mean 282.5 has been taken into account. In Figure 2 the trend is displayed graphically.

Table 3
The observed and computed proportion of "auxiliaries" in Rutherford's Nobel lecture

x	$p(Aux)$	$p(Aux)_{theor}$
1	0.0373	0.0693
2	0.0513	0.0794
3	0.0875	0.0908
4	0.0227	0.1037
5	0.0976	0.1182
6	0.1250	0.1344
7	0.2273	0.1524
8	0.2500	0.1944
9	0.3333	0.2185
10	0.1667	0.2446
11	0.1667	0.2728
13	0.4000	0.3030
14	0.2500	0.3349
16	0.5000	0.4032
20	0.5000	0.5489
60.5	0.9286	0.9979
282.5	1.0000	1.0000
$y = 1/(1 + 15.5663 \exp(-0.1471x))$, $R^2 = 0.94$, $IP = 18.66$		

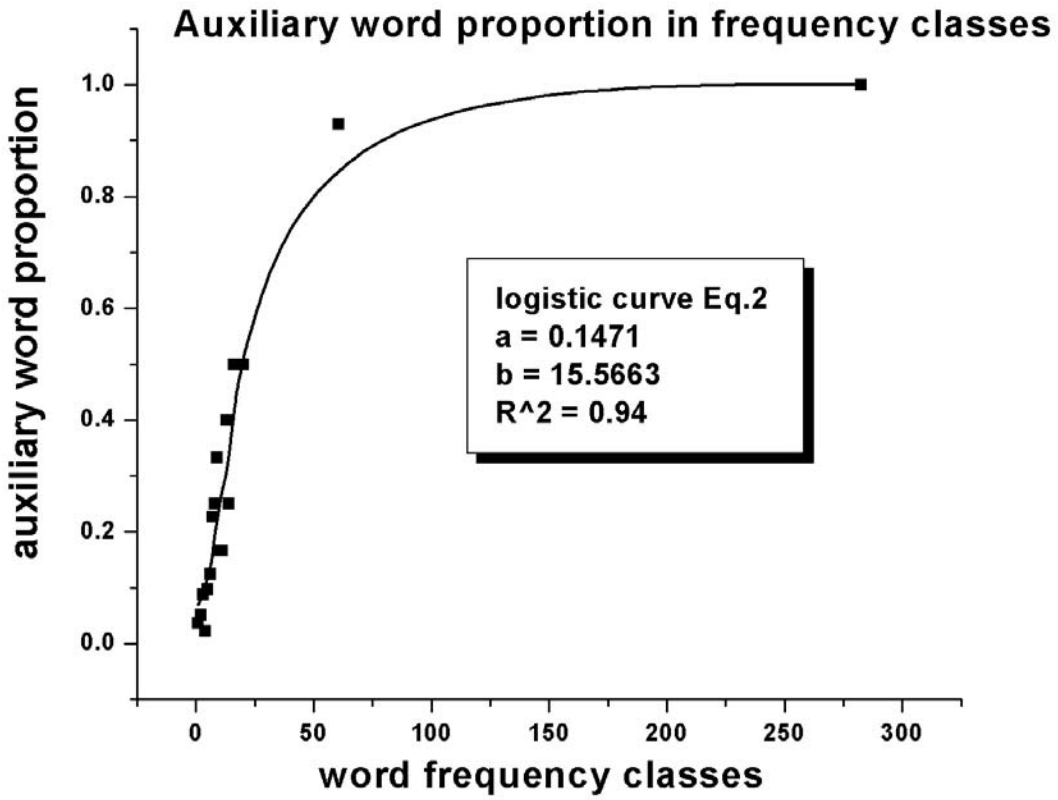

Figure 2. The dynamics of auxiliaries in an English text (Rutherford´s Nobel lecture)

The fit is very good, evidently we are dealing with a tendency that holds at least in some European languages. Of course, a number of further tests are necessary to corroborate this. Comparing the parameters *a* in German and English, we can preliminarily say that the greater *a*, the smaller is the analyticity (= the greater is syntheticity) in language; but this is a very preliminary statement requiring a lot of testing. We wish only to draw attention to this possibility.

3. The curves of word classes

One natural question which arises now is whether other word classes abide by this (or the opposite) trend; but we entrust this problem to the interested reader. Here we shall consider another question. As is generally known, the rank-frequency distributions of word classes follow the Zipf-law which – in the language of distributions – is represented by the (usually right-truncated) zeta distribution or some of its manifold generalisations and modifications (cf. Zörnig, Altmann 1995; Wimmer, Altmann 1999; Baayen 2001). Though the majority of these can be subsumed under a very general model (cf. Wimmer, Altmann 2005), we can consider the rank-frequency distribution of word forms from another point of view, namely as a mixture of word classes, each of which has its own distribution. The question is, what kind of distribution word classes follow, if any. The problem cannot be solved without first solving the linguistic problem of word classes. The PoS-taggers analysing word forms go to different depths and a false attribution can distort the form of the distribution. But even a "hand made" analysis involves decisions based on two hundred (or two thousand) years of discussion. Thus

we rely on taggers and perform two experiments. First, we consider the frequency spectrum of word classes (for nouns, verbs, adjectives and adverbs): a quite usual distribution in which some values of the variable are not realized. For the Rutherford text, we obtain the result in Table 4.

Table 4
Frequency spectra of word classes (Rutherford's Nobel lecture)

Nouns		Verbs		Adjectives		Adverbs	
x	g(x)	x	g(x)	x	g(x)	x	g(x)
1	205	1	114	1	105	1	38
2	85	2	40	2	30	2	28
3	31	3	28	3	13	3	11
4	21	4	9	4	7	4	4
5	16	5	7	5	11	5	3
6	13	6	3	6	2	6	1
7	7	7	4	7	2	7	3
8	8	8	4	8	4	8	2
9	6	9	1	9	1	11	1
10	1	10	3	12	1	12	1
11	4	12	3	16	1		
12	4	13	1	20	1		
13	2	14	1	22	1		
14	2	15	3				
15	1	17	1				
19	2	40	1				
28	1	42	1				
38	1	51	1				
44	1	85	1				
48	1						
51	1						
60	1						

It can easily be shown that any of the current distributions (Waring, Yule, Zipf-Mandelbrot, Good, zeta etc.) can be adequately fitted to these data. For the sake of illustration, we present in Table 5 the fit of the right truncated zeta distributions (Zipf) to the distribution of adjectives. The software inserts automatically zero for $g(x)$ if x is not realized, because the probability must be computed. It automatically pools frequency classes beginning from below to get the theoretical frequencies > 1. The chi-square test and the parameters of the zeta distribution are in the last row of the table. R is the truncation parameter at the right side of the distribution. In Figure 3 and 4 the fit is shown graphically. Similar results can be obtained also for nouns, verbs, adverbs and auxiliaries.

Table 5
Fitting the zeta distribution to adjectives in Rutherford´s Nobel lecture

x	g(x)	zeta
1	105	106,07
2	30	28,54
3	13	13,24
4	7	7,68
5	11	5,03
6	2	3,56
7	2	2,66
8	4	2,07
9	1	1,65
10	0	1,35
11	0	1,13
12	1	0,96
13	0	0,82
14	0	0,72
15	0	0,63
16	1	0,56
17	0	0,50
18	0	0,44
19	0	0,40
20	1	0,36
21	0	0,33
22	1	0,30
a = 1.89419; R = 22; DF =12; X^2 = 14,74; P = 0,26		

The right- truncated zeta distribution is defined as

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

where $T = \sum_{j=1}^{R} j^{-a}$.

This result is not surprising, but its consequences are rather strange. First, the resulting overall distribution looks (mathematically) like a superposition of identical distributions with differing parameters and different weights. But this is not true. The generating mechanism is the valency of the word classes. The appearance of a class in text leads to the appearance of another class which lies in the domain of its valency. Though the realization of a special class from this domain is controlled probabilistically (e.g. a noun admits an adjective but does not need to have it in any case, but on the contrary, an adjective requires a noun; a verb presupposes a noun or a pronoun but the pronoun can be an inflection, e.g. Latin "vocamus", etc.; this is realized differently in different languages), the associated classes follow the distribution of the main class.

Secondly, considering the frequencies of individual word classes as given in Table 4 and fitting a probability distribution to the data we have two problems: (i) the variable does not contain all the values; there are a number of *x*-values whose *f*(*x*) is zero. But we must compute the probability of these *x*, too, and use them for testing. (ii) Some expected frequencies are less

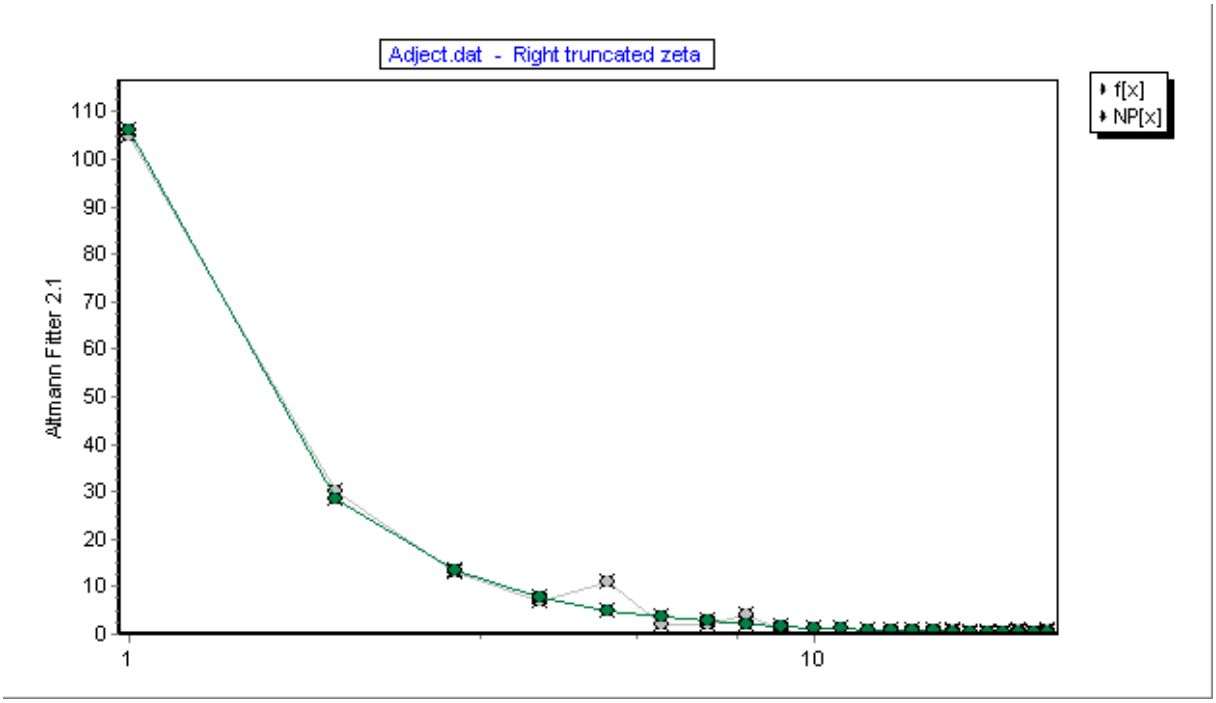

Figure 3. Fitting the right truncated zeta distribution to adjectives in Rutherford (half logarithmic presentation)

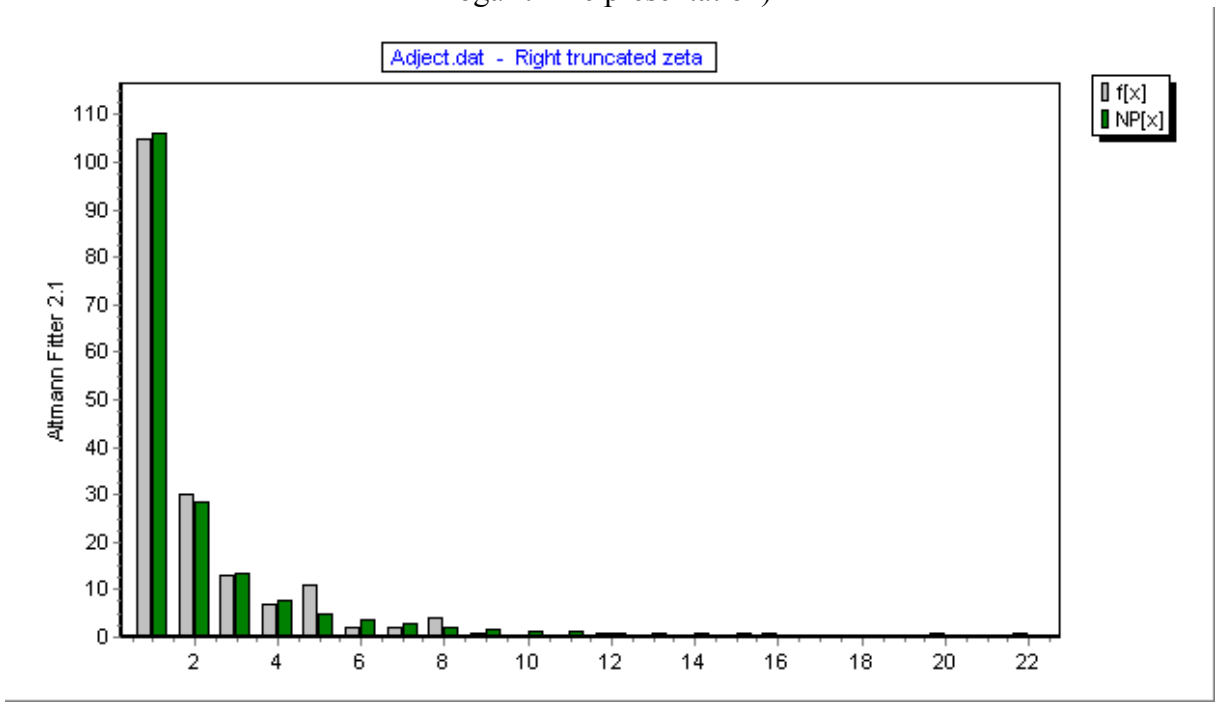

Figure 4. Fitting the right truncated zeta distribution to adjectives in Rutherford (bar presentation)

than 1, a circumstance which is not possible in reality unless it is zero, i.e. case (i). A part of the quantities in these classes has been, as a matter of fact, "shifted" to non-existing classes. In order to avoid these problems, we can set up a model of word-class distributions using an adequate curve. Using a curve or a sequence instead of a probability distribution, we can avoid the shortcomings above, but ignore the normalization. Real phenomena can be modelled by curves or sequences, probability distributions, discrete or continuous. There is no more truth in any of them; they are merely steps in the approximation of reality. Recalling

that Zipf used the zeta-function (not the zeta-distribution), we apply the same series, taking into account only the realized frequencies. Since the sequence alone converges to zero, we add a constant 1, i.e. we apply

(4) $y = bx^{-a} + 1$.

where y are the frequencies. From the data in Table 4 we obtain the results in Table 6.

Table 6
Fitting the modified zeta function to word-class distributions

Nouns			Verbs			Adjectives			Adverbs		
x	y	y_{theor}	x	y	y_{theor}	x	y	y_{theor}	x	y	y_{theor}
1	205	207.63	1	114	114.86	1	105	105.13	1	38	40.62
2	85	70.39	2	40	38.01	2	30	29.01	2	28	17.36
3	31	37.65	3	28	20.18	3	13	14.00	3	11	10.75
4	21	24.30	4	9	13.03	4	7	8.54	4	4	7.75
5	16	17.40	5	7	9.38	5	11	5.94	5	3	6.08
6	13	13.31	6	3	7.23	6	2	4.50	6	1	5.07
7	7	10.66	7	4	5.85	7	2	3.61	7	3	4.31
8	8	8.82	8	4	4.91	8	4	3.03	8	2	3.79
9	6	7.50	9	1	4.23	9	1	2.62	11	1	2.86
10	1	6.51	10	3	3.72	12	1	1.94	12	1	2.66
11	4	5.74	12	3	3.03	16	1	1.55			
12	4	5.13	13	1	2.78	20	1	1.36			
13	2	4.64	14	1	2.58	22	1	1.30			
14	2	4.24	15	3	2.41						
15	1	3.91	17	1	2.15						
19	2	3.00	40	1	1.29						
28	1	2.09	42	1	1.27						
38	1	1.67	51	1	1.19						
44	1	1.53	85	1	1.08						
48	1	1.47									
51	1	1.42									
60	1	1.33									

Nouns : $y = 206.6252x^{-1.5743} + 1$, $R^2 = 0.99$
Verbs : $y = 113.6624x^{-0.1563} + 1$, $R^2 = 0.99$
Adjectives : $y = 104.1267x^{-1.8942} + 1$, $R^2 = 0.996$
Adverbs : $y = 39.6188x^{-1.2761} + 1$, $R^2 = 0.89$

Note that for adjectives, both the zeta distribution and the above power function have the same parameter $a = 1.8942$. The above modified zeta-function for adjectives is shown in Figure 5.

We can conclude that not only the overall spectrum is zeta-like (Zipf-like) but also the individual main classes of words. Even mixing two classes (of the same text), e.g. verbs and adjectives, yields a zeta-like result. This picture can change if we take very long or very short texts. Since very long texts are necessarily mixtures whose parts abide by the same regime, however with different parameters, the mixture can cause model modifications. On the other hand, in very short texts the classes cannot take shape sufficiently. Nevertheless, in general,

both the overall spectrum and that of individual word classes abide by the power law or its modifications.

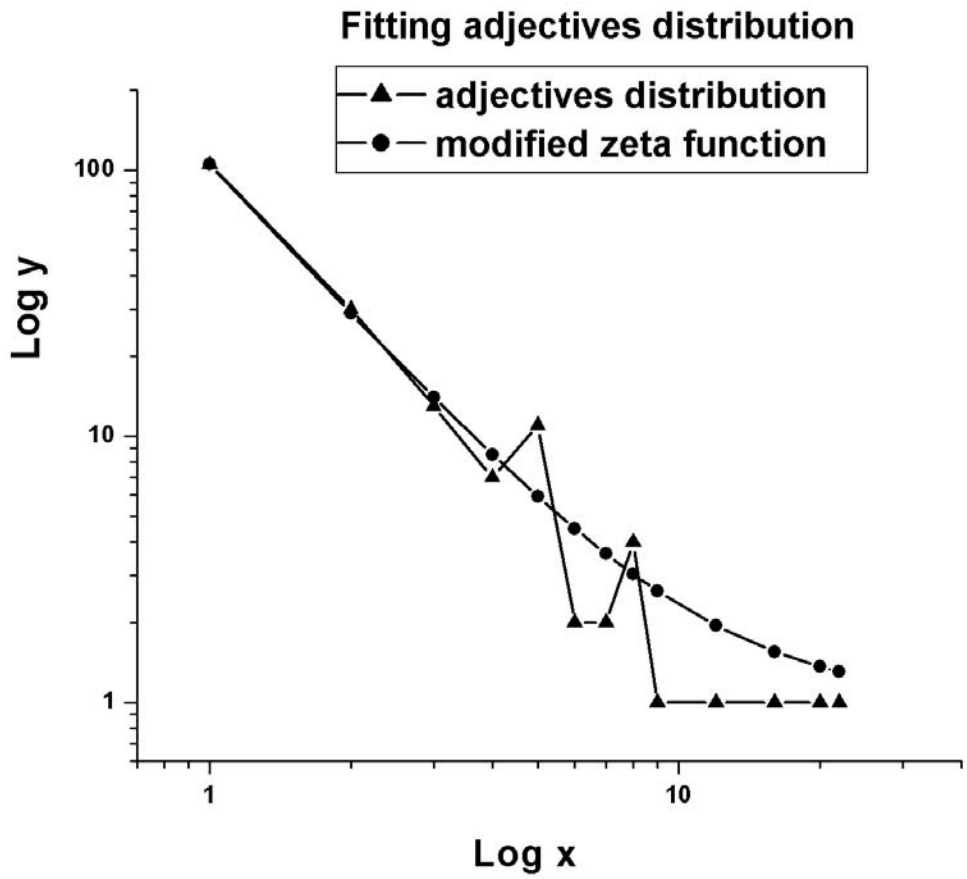

Figure 5. Fitting the modified zeta-function (4) to the distribution of adjectives

If we believe in the existence of laws in language, then from the above facts we can draw the conclusion that if an established word class in any language does not follow this law, it has been established incorrectly. This concerns even the well analysed languages. Ambiguities are mostly resolved syntactically. But even in that case many uncertainties remain which can be resolved using the frequency distribution.

Further research possibilities in this domain are: (i) Has parameter *a* of the power function (distribution or curve) characteristic values for individual word classes? (ii) Does the above regularity hold only for the main classes or for all? Which classes do not abide by this regularity? (iii) If the frequency spectrum deviates from the power law – and this has been shown in many cases – why does this happen, at which places, to what extent and in what direction? That is, we can begin to search for the boundary conditions of the rise of a word-frequency distribution.

4. Thematic concentration

The thematic concentration of a text can be evaluated intuitively, i.e. by reading it and expressing an expert opinion. Another possibility is to let a program search for key concepts in the text and to then try to interpret the results in terms of thematic concentration. Some tasks

of content analysis concern this aim. A third method is denotative analysis, also taking into account references and setting up a text-core (cf. Ziegler, Altmann 2002). A mechanization of this task is not yet fully developed. However, easy methods, which could not only help to find the core of the text but also measure its conspicuousness, are always highly desirable. In the following such a method will be proposed.

In a previous work (Popescu 2006) it has been shown that in the (monotonously decreasing) rank-frequency distribution of words or word forms there is a point in which $r = f(r)$, i.e. the rank of a word is equal to its frequency. This point is called h-point and it is the nearest point to the origin [0,0]. This point separates in a fuzzy way the great class of auxiliaries from autosemantic words. Of course, auxiliaries are found throughout the lower part (tail) of the distribution (below the *h*-point) but the point itself can be used for different purposes (cf. Popescu, Altmann 2006, 2007). Since we consider the pre-*h* domain as that of auxiliaries, autosemantics occurring in this domain must be extremely emphasized in the text, i.e. they must be part of the text theme, either its subject or its properties or activities. Sometimes they are proper names. This holds for any language, even extremely synthetic ones.

Consider the first thirty most frequent words in Rutherford´s Nobel lecture in Table 7. We

Table 7
The first thirty most frequent words in Rutherford´s Nobel lecture

r	f(r)	Word	r	f(r)	Word	r	f(r)	Word
1	466	the	11	60	it	21	39	atom
2	382	of	12	58	from	22	36	with
3	140	a	13	56	particles	23	30	for
4	121	that	14	53	helium	24	29	rays
5	116	and	15	51	is	25	28	an
6	113	in	16	45	particle	26	27	as
7	87	to	17	42	be	27	25	its
8	85	was	18	42	this	28	24	on
9	64	radium	19	41	at	29	24	or
10	63	by	20	40	were	30	23	radioactive

can see that the h-point is located at $r = 26$, or more exactly, at $h = 26.5$. There are only a few autosemantics with a rank smaller than *h*, namely *radium, particles, helium, particle, atom, rays*. From these words one can easily reconstruct what Rutherford spoke about. They build the primary theme of his lecture. In order to characterize quantitatively the concentration on these thematic words, we propose the following index of thematic concentration:

$$TC = \frac{2}{h}\sum_{r'=1}^{T}\frac{(h-r')f(r')}{(h-1)f(1)}. \qquad (5)$$

Here *h* is the *h*-point, $f(1)$ is the frequency of the most frequent word, r' are those ranks which point to thematic autosemantics, $f(r')$ is the frequency at these ranks and *T* is the number of those ranks. If there are no autosemantics in this domain, then $f(r')$ is in each case zero, *T* is zero, hence $TC = 0$. In the other extreme case – when all these words are autosemantics with frequency theoretically equal to $f(1)$ – we obtain $T = h$ and adding

$$\sum_{r=1}^{h}(h-r)=h(h)-\sum_{r=1}^{h}r = h^2 - \frac{h(h+1)}{2} = \frac{h(h-1)}{2}.$$

Dividing the sum by this constant we obtain formula (5) which is normalized and cannot surpass 1. Hence $TC \in <0, 1>$. Let us illustrate the computation using the above data from Rutherford. There are 6 autosemantics in the pre-h-domain, i.e. $T = 6$; $f(1) = 466$ and $h = 26$, hence

$$\frac{2}{h(h-1)f(1)} = \frac{2}{26(25)466} = 0.0000066028 = K$$

by which the sum will be multiplied. For the individual words we obtain:

word	rank	frequency	$(h - r')f(r')K$
radium	9	64	0.00718
particles	13	56	0.00481
helium	14	53	0.00420
particle	16	45	0.00297
atom	21	39	0.00129
rays	24	29	0.00038.

The sum of the last column yields $TC = 0.02083$. As can be seen, the realized non-zero values are of the order of ten of thousandth. Thus, defining the value of 1 *tcu* (thematic concentration unit) as a thematic concentration having $TC = 1/1000$, the thematic concentration of Rutherford's Nobel lecture has a value of 20.83 *tcu*. Needless to say, a lemmatised text would yield different values.

For comparative purposes we evaluated some other Nobel lectures and obtained the results as given in Table 8. The thematic words show automatically the field of the author.

Table 8
The TC-values of some Nobel lectures
(r_{min} is the smallest thematic rank in the pre-*h* domain)

Author	***h***	***f*(1)**	**pre-*h* words**	***TC***
Banting, F.G.	32	622	insulin, sugar, diet, patient, blood, carbohydrate	0.01692
Bellow, S.	26	297	art	0.00027
Buchanan,J.M.Jr.	23	366	political, politics, individual, individuals, rules	0.00985
Buck, P.	39	617	people, novel, Chinese, novels, China	0.01034
Feynman,R.P.	41	780	time, theory, quantum, electrodynamics	0.02222
Lewis, S.	25	237	American, America	0.00494
Mcleod, J.	24	460	insulin, sugar, pancreas, blood, symptoms	0.00604
Marshall,G.C.	19	229	peace	0.00506
Pauling, L.	28	546	nuclear, world, weapons, war, great, nations, human	0.01935
Russell, B.	29	342	power	0.00022
Rutherford, E.	26	466	radium, particles, helium, particle, atom, rays	0.02083

Multiplying *TC* by 1000 we obtain the following order (Table 9):

Table 9

Author	**Field**	***tcu***
Rutherford, E.	Chem	20.83
Pauling, L.	Peace	19.35
Banting, F.G.	Med	16.92
Buck, P.	Lit	10.34
Buchanan, J.M. Jr.	Econ	9.85
Mcleod, J.	Med	6.04
Marshall, G.C.	Peace	5.06
Lewis, S.	Lit	4.94
Feynman, R.P.	Phys	2.22
Bellow, S.	Lit	0.27
Russell, B.	Lit	0.22

In general, the representatives of "hard sciences" write texts with greatest thematic concentration. Linus Pauling was also an outstanding chemist, P. Buck and R. Feynman are "exceptions". But if we examine different genres, the picture will possibly be different. In any case art and social sciences will follow "hard science" and within art, poetry will have the smallest thematic concentration. Further research must be made on a very broad basis.

The above results lead automatically to the consideration of following possibilities: (i) Setting up a *post-h domain* consisting of ranks ($h,\ 2h - r_{min}$), r_{min} being the smallest thematic rank in the pre-*h* domain, e.g. with Rutherford $h = 26$, $r_{min} = 9$, from which $h - r_{min} = 17$, i.e. ($h,\ 2h - r_{min}$) = (26, 43) This domain can be called secondary thematic concentration domain. (ii) Using the pace of $h - r_{min}$ (which is 17 for Rutherford), one could partition the whole rank-frequency distribution in subsequent domains and study their contribution to thematic concentration. Preliminary computations have, however, shown that even the secondary domain contributes very little to the thematic concentration. Nevertheless, the series of domains could display some regularity concerning not only the theme of the text but also the dynamics of word classes. Here we shall not follow this possible direction of research.

Conclusions

The present article shows the first steps in scrutinizing the dynamics of word classes in text. Some trends are very regular and need a broad investigation in different languages. The whole rank-frequency domain can be partitioned in intervals based on the *h*-point and the behaviour of word classes can be studied within these intervals. The classes can be defined formally (e.g. length classes), grammatically (e.g. parts of speech) or semantically (e.g. words expressing activity). The dynamics can serve for the characterization of individual texts, of genres, epochs or even languages. The research can be performed with very simple mathematical apparatus.

References

Baayen, R.H. (2001). *Word frequency distributions*. Dordrecht: Kluwer**.**

Best, K.-H. (1994). Word class frequencies in contemporary German short prose texts. *Journal of Quantitative Linguistics 1, 144-147.*

Best, K.-H. (1997). Zur Wortartenhäufigkeit in Texten deutscher Kurzprosa der Gegenwart. *Glottometrika 16, 276-285.*

Best, K.-H. (2000). Verteilungen der Wortarten in Anzeigen. *Göttinger Beiträge zur Sprachwissenschaft 4, 37-51.*

Best, K.-H. (2001). Zur Gesetzmäßigkeit der Wortartenverteilungen in deutsche Pressetexten. *Glottometrics 1, 1-26.*

Hammerl, R. (1990). Untersuchungen zur Verteilung der Wortarten im Text. *Glottometrika 11, 142-156.*

Pawlowski, A. (2005). Modelling the sequential structures in text. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 738-750.* Berlin: de Gruyter.

Popescu, I.-I. (2006). The ranking by the weight of highly frequent words. In: Grzybek, P., Köhler, R. (eds.), *Exact methods in the study of language and texts: 553-562*. Berlin: Mouton-de Gruyter

Popescu, I.-I., Altmann, G. (2006). Some aspects of word frequencies. *Glottometrics 13, 23-46.*

Popescu, I.-I., Altmann, G. (2007). Some geometric properties of word frequency distributions. *Göttinger Beiträge zur Sprachwissenschaft (in print).*

Schweers, A., Zhu, J. (1991). Wortartenklassifizierung im Lateinischen, Deutschen und Chinesischen. In: Rothe, U. (ed.), *Diversification processes in language: grammar: 157-165.* Hagen: Rottmann.

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

Wimmer, G., Altmann, G. (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 191-208.* Berlin: de Gruyter.

Ziegler, A. (1998). Word class frequencies in Brazilian-Portuguese press texts. *Journal of Quantitative Linguistics 5, 269-280.*

Ziegler, A. (2001). Word class frequencies in Portuguese press texts. In: Uhlířová, L. et al. (eds.), *Text as a linguistic paradigm: levels, constituents, constructs. Festschrift in Honour of Luděk Hřebíček: 295-312.* Trier: WVT.

Ziegler, A., Altmann, G. (2002). *Denotative Textanalyse*. Wien: Praesens.

Ziegler, A., Best, K.H., Altmann, G. (2001). A contribution to text spectra. *Glottometrics 1, 97-108.*

Ziegler, A., Best, K.H., Altmann, G. (2002). Nominalstil. *ETC – Empirische Text- und Kulturforschung 2, 72-85.*

Zörnig, P., Altmann, G. (1995). Unified representation of Zipf distributions. *Computational Statistics & Data Analysis 19, 461-473.*