Printed in "Problems of General, Germanic and Slavic Linguistics"
Papers for 70th anniversary of Professor V. Levickij
Chernivtsi: Books - XXI, 2008, pp 472 - 480

Autosemantic compactness of texts

Ioan-Iovitz Popescu, Bucharest
Gabriel Altmann, Lüdenscheid

Abstract. In the article some new concepts of text building based on the h-point are introduced, namely the *thematic concentration* considering the proportion of autosemantics above the h-point, *autosemantic pace filling* expressing the proportions of autosemantics in h-intervals, and *autosemantic compactness* being a function of the parameters of the exponential curve representing the pace filling. Tests for comparing texts are proposed.

1. Each text has an infinite number of properties. From time to time one succeeds to coin conceptually a new property which may be intuitively cogent but its quantification and measurement (operationalization) may be a matter of long years of scientific discussion. A property can be operationalized in many different ways each of which can capture a special aspect of the given property. Remember the dozens of definitions of vocabulary richness based on the (intuitive) relation between V (vocabulary) and N (text length) or on the type-token ratio, and the rather technical definition using the Lorenz curve of deviations of the inverse cumulative rank-frequency distribution from the $P_{00} - P_{11}$ straight line which can be easily processed statistically (Popescu, Altmann 2006) and having a clear linguistic meaning, expressing the deviation from the maximal possible richness. Text properties are our conceptual constructs and are operationalized by non-unique definitions, e.g. such a simple property like "text length" is not "given" but defined in terms of numbers of phonemes, morphemes, syllables, word forms, lemmas, beats, phrases, clauses, sentences, hrebs (= denotative units), etc. And each of these units can in turn be defined in different ways as is known from 150 years of development of modern linguistics.

In this contribution we shall try to capture a phenomenon which can, perhaps, preliminarily be called autosemantic compactness of the text. Since its computation consists of a series of different steps, we first describe them individually.

2. One usually speaks about auto- and synsemantics, some authors distinguish autosemantics, functional words, auxiliaries, vicesemantics, synsemantics etc. It depends on the grammatical philosophy of the researcher or on the aim of the investigation which classes are established. For our purposes we define nouns as basic autosemantics. The basic autsemantics have two kinds of first order predicates, namely adjectives and verbs. All other word classes or parts of speech are either predicates of second or higher order, or auxiliaries, vicesemantics etc. In our analysis we shall be concerned only with "things" expressed as nouns and predicates of first order (adjectives and verbs). Everything else, also auxiliary and modal verbs will belong to the complement class. All this can easily be stated using the available PoS-taggers which can be downloaded from the Internet.

For a text under analysis a list of word forms or lemmas containing information about the affiliation of the word to a word class must be prepared. This goes also automatically using PoS-taggers from the Internet such as the CLAWS part-of-speech tagger for English at http://www.comp.lancs.ac.uk/ucrel/claws/. Some programs differentiate between uppercase and lowercase letters thus the text should be slightly (mechanically) pre-processed.

The list (sequence of word forms/lemmas) must be transformed in a frequency list and the counted entities should be ordered by decreasing frequency in order to obtain a rank-frequency distribution. (The argumentation with the frequency spectrum would be quite different). Word forms and lemmas could, perhaps, yield different results. Up to now there are no comparative studies. A good frequency counter can be found in the Internet, e.g. http://www.writewords.org.uk/word_count.asp.

In the rank-frequency distribution, $f = f(r)$, one finds a "nearest point" $r_{\min}$ which has the minimal distance $(r^2 + f^2)^{1/2}$ to the origin [0,0]. Thus, for instance, for a Zipf distribution $f(r) = c / r^a$ (with a and c constants), the rank of this point is given by $r_{\min} = a^{1/[2(1+a)]} c^{1/(1+a)}$. On the other side, not too far from it but, generally, distinct, there is another remarkable mathematical point (the so called "fixed point") for which $r = f(r)$, i.e. at which the rank equals the frequency. (i) Actually, in all our preceding papers, we meant and called this latter "the h-point". The rank of this point for the above Zipf distribution is given by $h_{Zipf} = c^{1/(1+a)}$, that is the two above points rank ratio is $r_{\min} / h_{Zipf} = a^{1/[2(1+a)]}$. As a numerical example, let us consider the Zipf fitting of Goethe's *Erlkoenig*, for which a = 0.6, hence $r_{\min} / h_{Zipf} = 0.8525$. Generally, only for symmetrical distributions, such as the particular Zipf distribution $f(r) = c / r$, that is for a = 1, the nearest point to the origin (0, 0) and the h-point would coincide at $r_{\min} = r_{Zipf} = c^{1/2}$. (ii) If there is not exactly such a $r = f(r)$ point, one can take the last r so that $r < f(r)$ and $r+1 > f(r+1)$, or (iii) one can take the mean of both, (iv) one takes the point $min[abs(r - f(r))]$. This is the meaning of what we called *h*-point (cf. Popescu 2006; Popescu, Altmann 2006, 2007; Popescu, Best, Altmann 2007). A further numerical example is shown in Table 1 using the first 30 most frequent word forms in Rutherford´s Nobel lecture. As can easily be seen, criterion (ii) is fulfilled at rank r = 26. The proportion of autosemantics above the *h*-point is called *thematic concentration* (c.f. Popescu, Best, Altmann 2007). For Rutherford we would obtain the autosemantics: *11 radium, 13 helium, 15 particles, 16 particle, 21 atom, 25 rays*, i.e. 6 autosemantics out of 26 pre-*h* words, resulting in a thematic concentration proportion of 6/26 = 0.23. Evidently, the result would slightly change if we consider lemmas.

Table 1
Ranking of the most frequent word forms
in E. Rutherford´s Nobel lecture

Rank	Frequency	Word form	PoS
1	464	THE	AT0
2	381	OF	PRF
3	140	A	AT0
4	121	THAT	CJT
5	116	AND	CJC
6	113	IN	PRP
7	87	TO	PRP
8	85	WAS	VBD
9	63	BY	PRP
10	60	IT	PNI
11	60	RADIUM	NP0
12	57	FROM	PRP

13	51	HELIUM	NN1
14	51	IS	VBZ
15	48	PARTICLES	NN2
16	44	PARTICLE	NN1
17	42	THIS	DT0
18	42	BE	VBI
19	41	AT	PRP
20	40	WERE	VBD
21	38	ATOM	NN1
22	36	WITH	PRP
23	29	FOR	PRP
24	28	AN	AT0
25	28	RAYS	NN2
26	27	AS	CJS
27	25	ITS	DPS
28	24	ON	PRP
29	23	OR	CJC
30	22	RADIOACTIVE	AJ0

The main purpose of the present paper is to bring arguments in favor of using the *h*-point also as a natural yardstick for the whole rank-frequency distribution. A simple application consists in measuring the stepwise crowding of autosemantics along the rank axis in fixed *h* steps, e.g. in Rutherford´s text in steps having a size of $h = 26$ successive ranks each. For this purpose, the rank axis should be divided in V/h intervals, where V is the text vocabulary, as illustrated in Table 2 for $V/h = 995$ (ranks)/26 (ranks/step) = 38.27 steps. The smallest crowding is 0, the maximal crowding is *h*. Evidently, it can be assumed that with increasing rank (and decreasing frequency) the number of autosemantics grows. If we partition the text in intervals of *h* ranks and state the number of autosemantics in each of them, we obtain a monotonously increasing function. Since the "steps" or h-intervals fill very quickly with autosemantics, the rate of change of crowding is the smaller the nearer is the step to its asymptote, say *a*. We assume a proportional relation and set

$$(1) \qquad \frac{dy}{dx} = k(a - y),$$

where *y* is the crowding of autosemantics, *x* is the step, *k* is a proportionality constant and *a* is the asymptote. Solving (1) and assuming $y(0) = 0$, we obtain

$$(2)\ y = a(1 - \exp(-kx)),$$

representing the increase of the number of autosemantics (their crowding) in subsequent *h*-intervals. Here *a* and *k* are parameters, *a* is the asymptote. In Fig. 1 one finds the general form of the autosemantics crowding function (2).

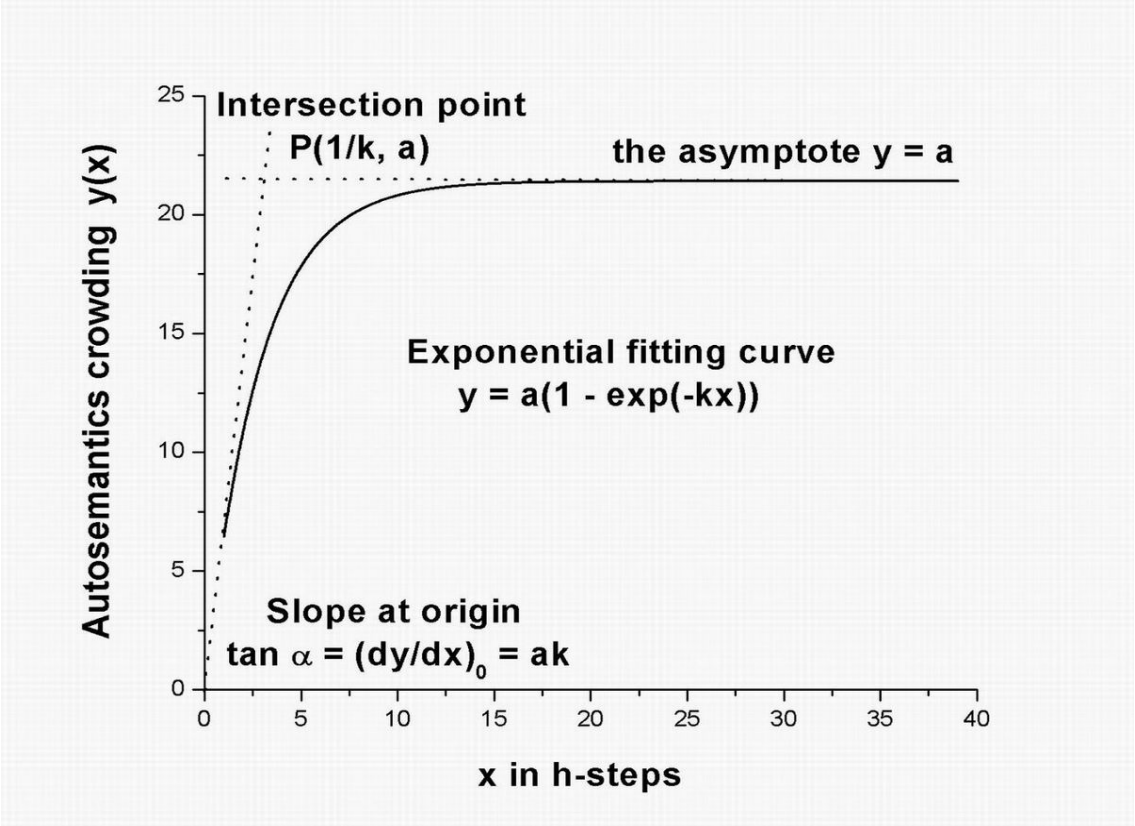


Figure 1. The exponential function as used to fit stepwise autosemantics filling of the rank-frequency distribution partitioned in *V/h* steps.

Generally, in our application, the parameter $a \leq h$ and the empirical points may attain *h* before the last *h*-interval but the fitted curve capturing the means of points need not. In Table 2 we present the empirical values of Rutherford´s text that are best fitted in Figure 2 by the following function

$$y = 21.40658(1 - \exp(-0.35932x))$$

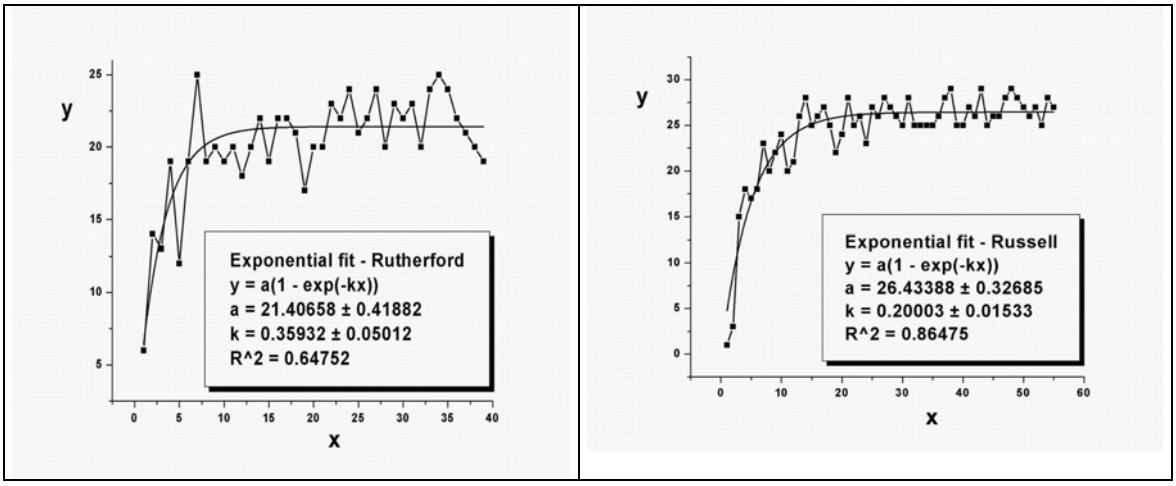
One can see that the curve does not reach $h = 26$ (its asymptote is $a = 21.407$) but it captures the increasing crowding of autosemantics in a satisfactory way.

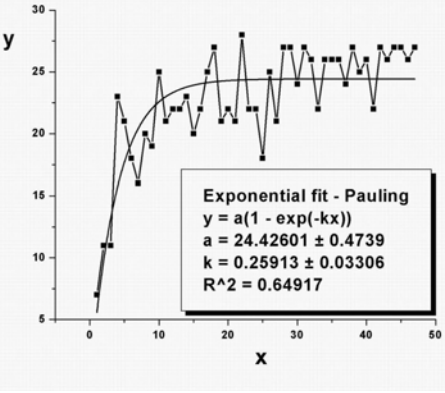
Table 2
Stepwise crowding of autosemantics in Rutherford´s Nobel lecture

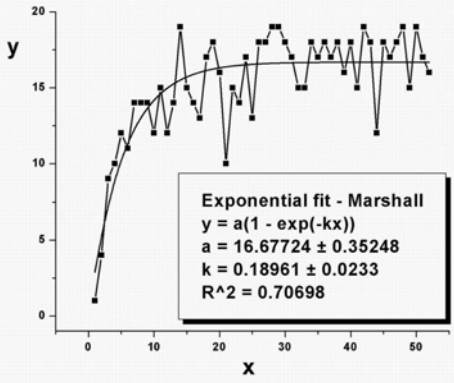
yardstick $h = 26$	re-ranking x paces	autosemantic pace filling y	yardstick $h = 26$	re-ranking x paces	autosemantic pace filling y
26	1	6	546	21	20
52	2	14	572	22	23
78	3	13	598	23	22
104	4	19	624	24	24
130	5	12	650	25	21
156	6	19	676	26	22

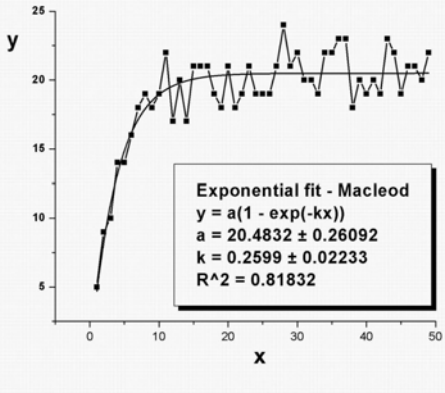
182	7	25	702	27	24
208	8	19	728	28	20
234	9	20	754	29	23
260	10	19	780	30	22
286	11	20	806	31	23
312	12	18	832	32	20
338	13	20	858	33	24
364	14	22	884	34	25
390	15	19	910	35	24
416	16	22	936	36	22
442	17	22	962	37	21
468	18	21	988	38	20
494	19	17	995	38.27*)	7
520	20	20	*) V/h = 995 (ranks)/26 (ranks/pace) = 38.27 paces		

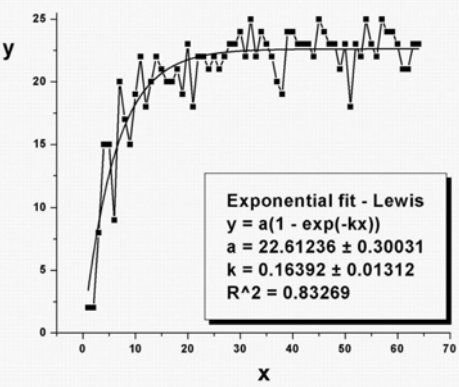
In Fig. 2 we present some graphs of different texts (Nobel lectures). Since the *h*-intervals are relatively small, the fluctuation of autosemantic crowding is relatively great, hence we obtain smaller determination coefficients but all F and t-tests are highly significant. For example the t-tests for the parameters in Rutherford yield t_a = 51.11, t_k = 7.17 and the F-test for regression yields F = 67.97, all very highly significant, though the determination coefficient is R^2 = 0.65 only. In general the exponential function and its common linguistic derivation seem to be a satisfactory solution. "Better" curves can, of course, be found but their embedding in linguistic model building would be rather difficult.

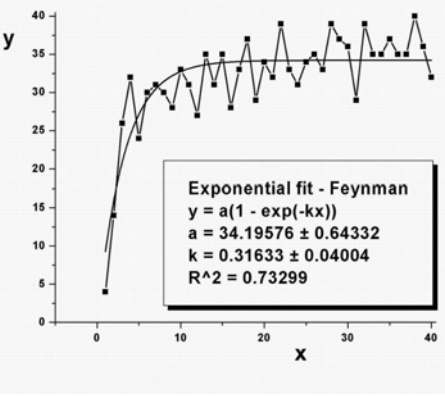


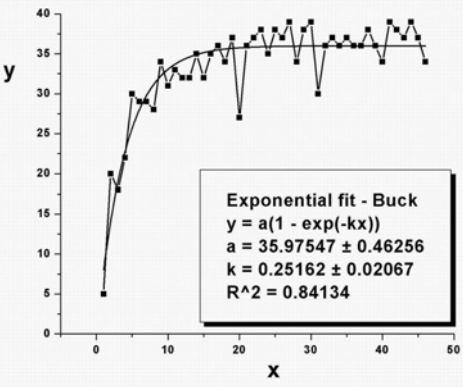

y
x
Exponential fit - Pauling
y = a(1 - exp(-kx))
a = 24.42601 ± 0.4739
k = 0.25913 ± 0.03306
R^2 = 0.64917


y
x
Exponential fit - Marshall
y = a(1 - exp(-kx))
a = 16.67724 ± 0.35248
k = 0.18961 ± 0.0233
R^2 = 0.70698


y
x
Exponential fit - Macleod
y = a(1 - exp(-kx))
a = 20.4832 ± 0.26092
k = 0.2599 ± 0.02233
R^2 = 0.81832


y
x
Exponential fit - Lewis
y = a(1 - exp(-kx))
a = 22.61236 ± 0.30031
k = 0.16392 ± 0.01312
R^2 = 0.83269


y
x
Exponential fit - Feynman
y = a(1 - exp(-kx))
a = 34.19576 ± 0.64332
k = 0.31633 ± 0.04004
R^2 = 0.73299


y
x
Exponential fit - Buck
y = a(1 - exp(-kx))
a = 35.97547 ± 0.46256
k = 0.25162 ± 0.02067
R^2 = 0.84134

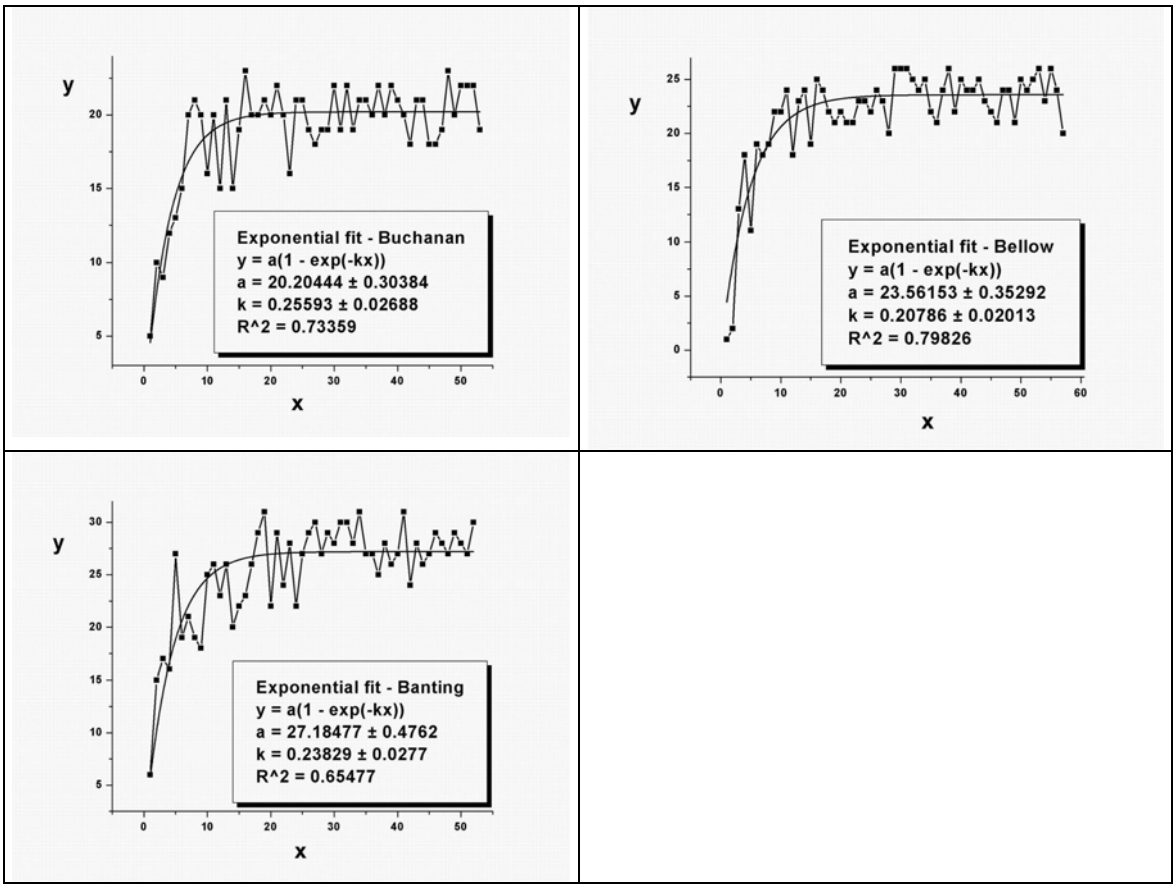


Figure 2. Graphs of autosemantic compactness of some Nobel lectures

The pictures of individual texts are very similar. It would be necessary to compare texts in different languages in order to obtain a more general idea of this regularity.

In Table 3 a survey of several Nobel lectures is presented..

Table 3
Autosemantics compactness and pace filling data of several Nobel lectures

Nobel	**Field**	***N***	***h***	***a***	***k***	**Compactness tan *α* = *ak***	**Pace filling *a/h***
Feynman	Phys	11265	41	34.196	0.316	10.806	0.834
Buck	Lit	9088	39	35.975	0.252	9.066	0.922
Rutherford	Chem	5083	26	21.407	0.359	7.685	0.823
Banting	Med	8193	32	27.185	0.238	6.470	0.850
Pauling	Peace	6246	28	24.426	0.259	6.326	0.872
Macleod	Med	4862	24	20.483	0.260	5.326	0.853
Russell	Lit	5701	29	26.434	0.200	5.287	0.912
Buchanan	Econ	4622	23	20.204	0.256	5.172	0.878
Bellow	Lit	4760	26	23.562	0.208	4.901	0.906
Lewis	Lit	5004	25	22.612	0.164	3.708	0.904
Marshall	Peace	3247	19	16.677	0.190	3.169	0.878

3. Some properties of the given curve can be interpreted textologically. As already mentioned above, the proportion of autosemantics in the first *h*-step can be considered *thematic concentration*. Since the second step consists also seldom of autosemantics only, it can be called *secondary thematic concentration*. Since both represent proportions, intertextual comparisons are easily possible using an asymptotic normal text.

The extent *a/h* to which the asymptote $y = a$ of curve (1) approaches the value of the pace *h* can be considered *autosemantic pace filling* (*APF*). The parameter *a* depends not only on the last pace but on all paces.

Consider for example the *APF* coefficients with Rutherford and Lewis. We obtain

	Rutherford	Lewis
h	26	25
a	21.407	22.612
APF	21.407/26 = 0.823	22.612/25 = 0.904

Since the standard deviation of any estimated parameter is automatically computed by the optimization program (see e.g. Fig. 2), the variance of *APF* can be obtained as $Var(APF) = Var(a)/h^2$, hence our test criterion will be

$$(3) \qquad z = \frac{APF_1 - APF_2}{\sqrt{Var(APF_1) + Var(APF_2)}} = \frac{APF_1 - APF_2}{\sqrt{\frac{Var(a_1)}{h_1^2} + \frac{Var(a_2)}{h_2^2}}}.$$

In our case we obtain

$$z = \frac{0.823 - 0.904}{\sqrt{\frac{0.41882^2}{26^2} + \frac{0.30031^2}{25^2}}} = 4.03$$

signalizing a highly significant difference between the *APF* of Rutherford and Lewis.

Consider now the properties of the curve in Figure 1. The text is autosemantically the more compact the steeper the slope of the curve. Since the tail of the rank-frequency distribution is always "almost full" of autosemantics, it is its beginning that contributes more to autosemantic compactness (and thematic concentration). Thus the slope of the curve at $x = 0$ can be considered a characteristic of the autosemantic compactness (*AC*). For curve (1) we obtain

$$(4) \qquad \frac{dy}{dx} = ak \exp(-kx)$$

and inserting $x = 0$ we get

$$(5) \qquad \tan \alpha = ak = AC$$

consisting of both parameters. The differences between the authors are here much greater, the range of analysed texts is from 3.2 to 10.8.

Again, we can set up an asymptotic test ignoring the covariances between the parameters. Since we need the variance of *AC*, we derive it using Taylor expansion and obtain

$$(6) \qquad V(AC) = \left(\frac{\partial AC}{\partial a}\right)^2 Var(a) + \left(\frac{\partial AC}{\partial k}\right)^2 Var(k).$$

The individual expressions yield

$$(7) \qquad V(AC) = k^2 Var(a) + a^2 Var(k)$$

a simple expression in which we use the estimated value of the parameter as its expectation. Hence our asymptotic and approximate criterion for measuring the difference between two indices of autosemantic compactness will be

$$(8) \qquad z = \frac{AC_1 - AC_2}{\sqrt{V(AC_1) + V(AC_2)}}.$$

Using the data from Rutherford and Lewis we have

	Rutherford	Lewis
a	21.407	22.612
k	0.359	0.164
Var(*a*)	0.419^2	0.300^2
Var(*k*)	0.050^2	0.013^2.

Inserting these values in (7) we obtain for Rutherford

$$V(AC_{Rutherford}) = 0.359^2 0.419^2 + 21.407^2 0.050^2 = 1.16728$$

For Lewis we obtain

$$V(AC_{Lewis}) = 0.08883$$

Hence comparing Rutherford and Lewis we obtain

$$z = \frac{7.685 - 3.708}{\sqrt{1.16728. + 0.08883}} = 4.457 .$$

The difference is again highly significant. The adding of covariances rendered the difference somewhat smaller.

4. As shown in Fig. 1, the first derivation of the curve in point $x = 0$ yields the tangent of the curve, i.e. its slope. The straight line in this point is $y = akx$. On the other hand, the asymptote of the curve is $y = a$, hence the crossing point is P[1/*k, a*]. Thus the autosemantic compactness (by definition $\tan \alpha = ak$) varies proportionally with *k* and inverse proportionally with $x(P) = 1/k$. This point characterizes the autosemantic construction of the text and can be

used e.g. for performing discriminant or other taxonomic analyses. In Table 4 one finds these coordinates for all texts and in Figure 3 the points with names of authors are presented.

Table 4
Autosemantic coordinates of individual texts

Nobel	**Field**	***N***	***h***	***1/k***	***a***
Rutherford	Chem	5083	26	2,786	21,407
Feynman	Phys	11265	41	3,165	34,196
Macleod	Med	4862	24	3,846	20,483
Pauling	Peace	6246	28	3,861	24,426
Buchanan	Econ	4622	23	3,906	20,204
Buck	Lit	9088	39	3,968	35,975
Banting	Med	8193	32	4,202	27,185
Bellow	Lit	4760	26	4,808	23,562
Russell	Lit	5701	29	5,000	26,434
Marshall	Peace	3247	19	5,263	16,677
Lewis	Lit	5004	25	6,098	22,612

The number of texts is too small to ascertain a dependence of the coordinate points on text length *N*. The dispersion is preliminarily too great.

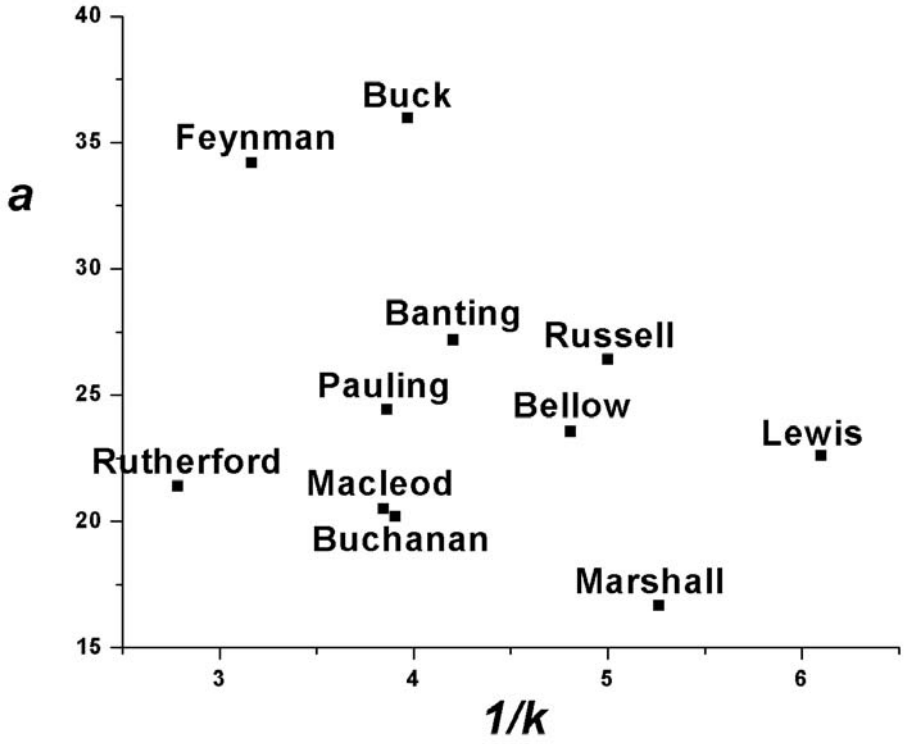


Figure 3. Autosemantic coordinates of individual texts

5. Conclusions

The dynamics of filling a text with autosemantics of first order (nouns, verbs, adjectives) can be quantified in different ways. Our aim was only to show some methods, operationalizations and testing possibilities. If texts in other languages could corroborate these results, one could tend to accept the above simple exponential dependence of autosemantic filling as a candidate for a text law. However, using one language only, the corroboration is very weak. In any case *a/h* seems not to be dependent of *N*.

On the other hand, the partitioning of words in autosemantics and the rest is only one of the great number of possibilities for which we do not even have concepts. Nevertheless, the above methods could be helpful in discovering finer aspects of word frequencies. One further aspect of autosemantics is their coincidence in texts giving rise to networks whose properties can be studied by the respective methods of graph theory.

References

Popescu, I.-Iovitz (2006). Text ranking by the weight of highly frequent words. In: Grzybek, Peter, Köhler, Reinhard (eds.), *Exact methods in the study of language and text: 553-562.* Berlin/New York: de Gruyter.

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

Popescu, I.-Iovitz; Altmann, Gabriel (2007). Some geometric properties of word frequency distributions. *Göttinger Beiträge zur Sprachwissenshcaft 13, 87-98.*

Popescu, I.-Iovitz; Best, Karl-Heinz; Altmann, Gabriel (2007). On the dynamics of word classes in text. *Glottometrics 14, 61-74.*