Glottometrics 15, 2007, 45-52

Confidence intervals and tests for the h-point and related text characteristics

Ján Mačutek[1] [2]*, Bratislava*
Ioan-Iovitz Popescu, Bucharest
Gabriel Altmann, Lüdenscheid

Abstract. Confidence intervals and tests for recently introduced text characteristics (the *h*-point and its relatives) are derived.

Keywords: text analysis, h-point, a- indicator

1. Introduction

The *h*-point was suggested by Hirsch (2005) as an index of research productivity (mainly) in physics. Popescu (2007) uses it in linguistics as a point which separates highly frequent synsemantic (or auxiliary) words from autosemantic words with lower frequencies. Popescu and Altmann (2006) introduce other three text characteristics – the *k*-point, the *m*-point and the *n*-point, which are modifications or analogies of the *h*-point applied to the frequency spectrum, the cumulative distribution or the reverse order rank-frequency distribution. All four of them can be used to measure vocabulary richness of texts, text coverage, text compactness, analytism and synthetism of language, and so on.

Synsemantic words are usually concentrated in the first classes of the rank-frequency distribution, while much more numerous autosemantic words tend to have significantly lower frequencies. However, there is no sharp boundary separating these two branches; in texts there are a few very frequent autosemantics (they build the theme of the text: see Popescu, Best and Altmann 2007), and/or some synsemantics with low frequencies which may have synonyms used alternately. Therefore we derive confidence intervals for the above mentioned characteristic points. The intervals should cover the area where synsemantics and autosemantics are mixed.

2. Confidence intervals

The *h*-point is an extension of the mathematical fixed point to the actual discrete rank-frequency distribution, $f = f(r)$ (of words in our case), that is by definition is the point $(r, f(r))$ where $f(h) = h$ (if such a point does not exist in the actual distribution, one takes that r whose absolute difference to $f(r)$ is minimum). Table 1 (see below) contains rank-frequency distribution of word forms in Goethe's poem "Erlkönig". The table is presented in Popescu and Altmann (2006).

[1] Address correspondence to: jmacutek@yahoo.com
[2] Supported by research grant VEGA 1/3016/06.

Table 1
Erlkönig

Rank r	Frequency f(r)	Cumulative frequency cf(r)	Relative cumulative frequency $F(r)$	Rank r	Frequency f(r)	Cumulative frequency cf(r)	Relative cumulative frequency F(r)
1	11	11	0.0489	21	3	104	0.4622
2	9	20	0.0889	22	2	106	0.4711
3	9	29	0.1289	23	2	108	0.4800
4	7	36	0.1600	24	2	110	0.4889
5	6	42	0.1867	25	2	112	0.4978
6	***6***	48	0.2133	26	2	114	0.5067
7	5	53	0.2356	27	2	116	0.5156
8	5	58	0.2578	28	2	118	0.5244
9	4	62	0.2756	29	2	120	0.5333
10	4	66	0.2933	30	2	122	0.5422
11	4	70	0.3111	31	2	124	0.5511
12	4	74	0.3289	32	2	126	0.5600
13	4	78	0.3467	33	2	128	0.5689
14	4	82	0.3644	34	2	130	0.5778
15	4	86	0.3822	35	2	132	0.5867
16	3	89	0.3956	36	2	134	0.5956
17	3	92	0.4089	37	2	136	0.6044
18	3	95	0.4222	38	2	138	0.6133
19	3	98	0.4356	39	2	140	0.6222
20	3	101	0.4489	40-124*	1	225	1

* The ranks 40 to 124 have frequency 1

It is easy to see that the *h*-point is 6, as $f(6) = 6$. We have $N = 225$ and $F(h) = 0.2133$.

Now, let *h* be the *h*-point, cp_h the cumulative probability at *h* and X_h the number of values which are less or equal to *h*. X_h can attain the values $0,1,2,\ldots,N$ with the probabilities which can be derived from (and explained by) an urn scheme consideration. Suppose we randomly place *N* balls into *V* urns labeled $1,2,\ldots,V$. Divide the urns into two groups – the "synsemantic group" consists of the urns $1,2,\ldots,h$, while the urns $h+1, h+2,\ldots,V$ belong to the "autosemantic group". We do not know (and do not need) the probabilities that a ball will be put into a particular urn. All we need is the probability that a ball will be placed into the "synsemantic group" of urns, which is the sum of probabilities of all the urns from that group (or, in other words, it the cumulative probability at *h*, i.e., cp_h). The probability of putting a ball into the "autosemantic group" of urns is, of course, uniquely determined by the previous one; it is equal to $1 - cp_h$.

$X_h = 0$ (i.e., all balls are in the "autosemantic group" of urns) with the probability

$$P(X_h = 0) = (1 - cp_h)^N,$$

$X_h = 1$ (i.e., one ball is in the "synsematic group", all the other balls are in the "autosemantic group") with the probability

$$P(X_h = 2) = \binom{N}{2} cp_h^2 (1 - cp_h)^{N-2}, \text{ etc;}$$

in general

$$P(X_h = r) = \binom{N}{r} cp_h^r (1 - cp_h)^{N-r}, \; r = 0,1,2,\ldots,N .$$

Hence, X_h has the binomial distribution with the parameters N and cp_h. We note that in general the considered urn scheme is not a non-increasing distribution and the confidence interval is approximate only.

Denote $c\hat{p}_h$ the estimation of cp_h, i.e., $c\hat{p}_h$ is the relative cumulative frequency at h. The binomial distribution can be approximated by the normal distribution. In the next step we obtain the confidence intervals for cp_h:

$$P\left(c\hat{p}_h - u_{1-\frac{\alpha}{2}} \sqrt{\frac{c\hat{p}_h(1-c\hat{p}_h)}{N}} \le cp_h \le c\hat{p}_h + u_{1-\frac{\alpha}{2}} \sqrt{\frac{c\hat{p}_h(1-c\hat{p}_h)}{N}} \right) = 1-\alpha, \qquad (1)$$

where α is the significance level and $u_{1-\alpha/2}$ is the quantile of the standard normal distribution which can be found in any statistical tables or statistical software. The probability remains unchanged if we multiply in (1) all three expressions in the parentheses by N:

$$P\left(Nc\hat{p}_h - u_{1-\frac{\alpha}{2}} \sqrt{Nc\hat{p}_h(1-c\hat{p}_h)} \le Ncp_h \le Nc\hat{p}_h + u_{1-\frac{\alpha}{2}} \sqrt{Nc\hat{p}_h(1-c\hat{p}_h)} \right) = 1\text{-}\alpha \quad (2)$$

We have obtained the confidence interval for cumulative frequencies at the h-point. The interval, where the cumulative frequencies from (2) are attained, is the confidence interval for the h-point. If the cumulative frequencies do not attain exactly the values equal to the confidence interval limits, we suggest taking the highest frequency below the lower interval limit, and the lowest frequency above the upper interval limit.

In the "Erlkönig" we have $c\hat{p} = 0.2133$ and N = 225. For $\alpha = 0.05$ the interval (2) (i.e., the confidence interval for cumulative probabilities) yields (35.95, 60.04). We have $cf(3)$ = 29, $cf(4)$ = 36 and $cf(8)$ = 58, $cf(9)$ = 62. Hence, [3,9] is at least 95% confidence interval for the h-point.

Confidence intervals for the k-point, m-point and n-point (all of them defined in Popescu and Altmann 2006) can be constructed in the same way, using, of course, the respective cumulative probabilities and cumulative frequencies.

3. Tests

3.1. Test for cumulative probabilities corresponding to *h*-points

The approach from the previous section can also be applied to obtain a test for comparing cumulative probabilities corresponding to h-points in two different texts.

Consider two texts. Let h_1, h_2 denote their h-points, cp_{h1}, cp_{h2} the cumulative probabilities

at h_1, h_2 (with $c\hat{p}_{h1}, c\hat{p}_{h2}$ being their estimations) and N_1, N_2 the numbers of word forms or lemmas in the texts, respectively. The statistic

$$U = \frac{\hat{cp}_{h1} - \hat{cp}_{h2}}{\sqrt{\frac{\hat{cp}_{h1}\left(1-\hat{cp}_{h1}\right)}{N_1} + \frac{\hat{cp}_{h2}\left(1-\hat{cp}_{h2}\right)}{N_2}}} \tag{3}$$

has approximately the standard normal distribution. Hence, in terms of corresponding cumulative probabilities, two *h*-points are significantly different if $|U| > u_{1-\alpha/2}$. Recall once more that (3) is a test for comparing cumulative probabilities corresponding to the *h*-points, not for comparing the *h*-points themselves. In other words, (3) can be used to test whether the ratios

$$\frac{number\ of\ word\ forms\ (lemmas)\ with\ frequencies\ higher\ than\ h}{number\ of\ all\ word\ forms\ (lemmas)}$$

in the texts under consideration are significantly different.

As an example, we test the difference of cumulative probabilities corresponding to the *h*-points in two poems – Goethe's "Erlkönig" (see Table 1 above) and Moericke's "Peregrina" (the rank-frequency distribution of word forms in which is presented in Table 2).

Table 2
Peregrina

Rank r	Frequency f(r)	Cumulative frequency cf(r)	Relative cumulative frequency F(r)	Rank r	Frequency f(r)	Cumulative frequency cf(r)	Relative cumulative frequency F(r)
1	16	16	0.0270	38	2	218	0.3676
2	16	32	0.0540	39	2	220	0.3710
3	12	44	0.0742	40	2	222	0.3744
4	12	56	0.0944	41	2	224	0.3777
5	11	67	0.1130	42	2	226	0.3811
6	10	77	0.1298	43	2	228	0.3845
7	8	85	0.1433	44	2	230	0.3879
8	***8***	93	0.1568	45	2	232	0.3912
9	7	100	0.1686	46	2	234	0.3946
10	7	107	0.1804	47	2	236	0.3980
11	6	113	0.1906	48	2	238	0.4013
12	6	119	0.2007	49	2	240	0.4047
13	6	125	0.2108	50	2	242	0.4081
14	6	131	0.2209	51	2	244	0.4115
15	5	136	0.2293	52	2	246	0.4148
16	5	141	0.2378	53	2	248	0.4182
17	5	146	0.2462	54	2	250	0.4216
18	5	151	0.2546	55	2	252	0.4250
19	5	156	0.2631	56	2	254	0.4283

20	5	161	0.2715	57	2	256	0.4317
21	4	165	0.2782	58	2	258	0.4351
22	4	169	0.2850	59	2	260	0.4384
23	4	173	0.2917	60	2	262	0.4418
24	4	177	0.2985	61	2	264	0.4452
25	4	181	0.3052	62	2	266	0.4486
26	3	184	0.3103	63	2	268	0.4519
27	3	187	0.3153	64	2	270	0.4553
28	3	190	0.3204	65	2	272	0.4287
29	3	193	0.3255	66	2	274	0.4621
30	3	196	0.3305	67	2	276	0.4654
31	3	199	0.3356	68	2	278	0.4688
32	3	202	0.3406	69	2	280	0.4722
33	3	205	0.3457	70	2	282	0.4755
34	3	208	0.3508	71	2	284	0.4789
35	3	211	0.3558	72	2	286	0.4823
36	3	214	0.3609	73	2	288	0.4857
37	2	216	0.3642	74-378*	1	593	1

* The ranks 74 to 378 have frequency 1

We have $h_1 = 6,\ c\hat{p}_{h1} = 0.2133,\ N_1 = 225$ (Erlkönig) and $h_2 = 8,\ c\hat{p}_{h2} = 0.1568,\ N_2 = 593$ (Peregrina). The test (3) yields

$$U = \frac{0.2133 - 0.1568}{\sqrt{\frac{0.2133(1-0.2133)}{225} + \frac{0.1568(1-0.1568)}{593}}} = 1.8153$$

which means that for $\alpha = 0.05$ we do not reject the hypotheses that the cumulative probabilities corresponding to the *h*-points in these poems are equal ($u_{0.975} = 1.96$).

3.2. Test for *a*-indices

The relationship between *h* and *N* can be expressed by a simple equation $N = ah^2$ (suggested by Hirsch 2005, mentioned also in Popescu and Altmann 2006). In fact, the quantity

$$a = \frac{N}{h^2} \tag{4}$$

does not depend any more on *N*, as can be shown using about 200 texts from 20 languages. On the contrary, index *a* is an indicator of language analyticity. The smaller *a* is, the more analytic the language is; the greater *a* is, the more synthetic the language is. As shown in Table 3, this statement can be corroborated empirically using 20 languages.

Table 3
Mean values of *a* in texts of 20 languages
(from Popescu et al. 2007)

Language	Mean a	Language	Mean a
Samoan	4.56	Italian	8.41
Rarotongan	5.02	Romanian	9.15
Hawaiian	5.37	Slovenian	9.19
Maori	5.53	Indonesian	9.58
Lakota	5.69	Russian	10.10
Marquesan	5.69	Czech	10.33
Tagalog	7.24	Marathi	11.82
English	7.65	Kannada	16.58
Bulgarian	7.81	Hungarian	18.02
German	8.39	Latin	19.56

As the index *a* is independent of *N*, it can be used to compare different texts (of different lengths). We need the variances of *a*-indices to construct the test; hence first we look for the distributions of the *h*-points. As a theoretical formula is not known, and all attempts to approximate it failed, a simulation study was used.

It was shown that word rank-frequency distributions in almost all texts can be modeled by the right truncated zeta distribution ($P_x = cx^{-a}$, $x = 1,2,\ldots,V$, with *c* being the normalization constant, cf. Wimmer and Altmann 1999, pp. 577-578), and its parameter can be easily estimated by Altmann-Fitter or other software.

The simulation study can be described as follows ("Erlkönig" being an example again). We generate 225 random numbers (there are 225 words in "Erlkönig") from the right truncated zeta distribution with the parameter 0.6007 (for this parameter value the best fit is obtained) and we find the *h*-point for their rank-frequency distribution. The random numbers generation is repeated 100-times (i.e., we have 100 *h*-points from the samples with the same size and with the same distribution as word frequencies in "Erlkönig"). The *a*-indices for these *h*-points are computed and their variance is found. Finally, the process was repeated 10-times, i.e., 10 variance values (each of them being a variance of 100 *a*-indices) were obtained. We take their mean as the variation of the *a*-index.

The above mentioned numbers of generations may be considered too low, but they require quite a lot of time and our aim is to present the method only.[3]

Now we can test the difference between the *a*-indices in two texts. Denote them a_1, a_2. The statistic

$$U_a = \frac{a_1 - a_2}{\sqrt{Var(a_1) + Var(a_2)}} \tag{5}$$

has, again, approximately the standard normal distribution, i.e., the difference between a_1 and a_2 is significant if $|U_a| > u_{1-\alpha/2}$.

[3] A short simulation program written in *R* can be sent upon request (jmacutek@yahoo.com).

In our texts by Goethe and Moericke we have

$$a_1 = \frac{N_1}{h_1^2} = \frac{225}{6^2} = 6.25 \text{ (Erlkönig)},$$

$$a_2 = \frac{N_2}{h_2^2} = \frac{593}{8^2} = 9.2656 \text{ (Peregrina)}.$$

On the other side, by the above simulation we obtained $Var(a_1) = 48.82$ for "Erlkönig" and $Var(a_2) = 99.05$ for "Peregrina", hence we finally have

$$U_a = \frac{6.25 - 9.2656}{\sqrt{48.82 + 99.05}} = -0.248,$$

which means that for $\alpha = 0.05$ the *a*-indices for "Erlkönig" and "Peregrina" are not significantly different.

4. Examples

In order to check the intralinguistic and extralinguistic differentiation of texts we performed the test for differences of cumulative probabilities corresponding to the *h*-points on 13 texts (by Goethe in German and Eminescu in Romanian, Table 4). The list of texts is given in the Appendix. The matrix is antisymmetric, i.e., the number in the *i*-th row and *j*-th column and the number in the *j*-th row and *i*-th column have the same absolute values but opposite signs. The critical value is ±1.96.

Table 4
U-test for the difference between cumulative probabilities
corresponding to *h*-points

	G 05	G 09	G 10	G 11	G 12	G 14	G 17	R 01	R 02	R 03	R 04	R 05	R 06
G 05	0	0.60	1.84	**2.64**	0.91	1.09	0.43	0.02	-0.38	0.33	0.86	0.20	0.28
G 09	-0.60	0	1.33	**2.16**	0.46	0.69	-0.01	-0.73	**-2.12**	0.37	**2.17**	-0.06	0.19
G 10	-1.84	-1.33	0	0.77	-0.58	-0.26	-0.99	**-2.24**	**-3.57**	-1.17	0.49	-1.51	-1.17
G 11	**-2.64**	**-2.16**	-0.77	0	-1.21	-0.83	-1.58	**-3.25**	**-4.63**	**-2.11**	-0.44	**-2.43**	**-2.01**
G 12	-0.91	-0.46	0.58	1.21	0	0.23	-0.38	-1.01	**-1.97**	-0.24	1.02	-0.53	-0.32
G 14	-1.09	-0.69	0.26	0.83	-0.23	0	-0.58	-1.19	**-2.03**	-0.50	0.61	-0.76	-0.56
G 17	-0.43	0.01	0.99	1.58	0.38	0.58	0	-0.46	-1.34	0.26	1.44	-0.03	0.15
R 01	-0.02	0.73	**2.24**	**3.25**	1.01	1.19	0.46	0	-1.86	1.38	**3.81**	0.78	0.98
R 02	0.38	**2.12**	**3.57**	**4.63**	**1.97**	**2.03**	1.34	1.86	0	**3.17**	**5.80**	**2.41**	**2.42**
R 03	-0.33	-0.37	1.17	**2.11**	0.24	0.50	-0.26	-1.38	**-3.17**	0	**2.22**	-0.49	-0.15
R 04	-0.86	**-2.17**	-0.49	0.44	-1.02	-0.61	-1.44	**-3.81**	**-5.80**	**-2.22**	0	**-2.59**	**-2.00**
R 05	-0.20	0.06	1.51	**2.43**	0.53	0.76	0.03	-0.78	**-2.41**	0.49	**2.59**	0	0.27
R 06	-0.28	-0.19	1.17	**2.01**	0.32	0.56	-0.15	-0.98	**-2.42**	0.15	**2.00**	-0.27	0

As can be seen, not only different authors but also different works of the same author may display significant differences, even in the same genre. Hence the *h*-point and the derived indicator *a* can be considered text-dependent characteristics.

References

Altmann-Fitter (1997). *Iterative Anpassung diskreter Wahrscheinlichkeitsverteilungen.* Lüdenscheid: RAM-Verlag.

Hirsch, J.E. (2005). An index to quantify an individual's research output. *Proceedings of the National Academy of Sciences of the USA 102, 16569-16572.*

Popescu, I.-I. (2007). Text ranking by the weight of highly frequent words. In: Grzybek, P., Köhler, R. (eds), *Exact methods in study of language and text, 553-562,* Berlin / New York: de Gruyter.

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

Popescu, I.-I., Best, K.H., Altmann, G. (2007). On the dynamics of word classes in texts. *Glottometrics 14, 58-71.*

Popescu, I.-I., Vidya, M.N., Uhlířová, L., Pustet, R., Mehler, A., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B.D., Grzybek, P., Altmann, G. (2007). *Word frequency studie*s. (In press)

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

Appendix : Texts used

G 05:	Goethe, J.W.v.	Der Gott und die Bajadere
G 09:	Goethe, J.W.v	Elegie 19
G 10:	Goethe, J.W.v	Elegie 13
G 11:	Goethe, J.W.v	Elegie 15
G 12:	Goethe, J.W.v	Elegie 2
G 14:	Goethe, J.W.v	Elegie 5
G 15:	Moericke, E.	Peregrina
G 17:	Goethe, J.W.v	Der Erlkönig
R 01:	Eminescu, M.	Luceafarul - Lucifer
R 02:	Eminescu, M.	Scrisoarea III - Satire III
R 03:	Eminescu, M.	Scrisoarea IV - Satire IV
R 04:	Eminescu, M.	Scrisoarea I - Satire I
R 05:	Eminescu, M.	Scrisoarea V - Satire V
R 06:	Eminescu, M.	Scrisoarea II - Satire II