Analysing h-point in lemmatised and non-lemmatised texts

Emmerich Kelih, Andrij Rovenchak, Solomija Buk

1. Introduction

In quantitative text analysis, the h-point plays an important role and is an important requirement for the calculation of further statistical stylistic and typological parameters. The h-point is a fixed point of a rank frequency distribution and is mainly used in lexical and word frequency studies. From a linguistic point of view the h-point is mainly understood as a fuzzy border between autosemantic and synsemantic word forms (cf. Popescu et al. 2009: 23). This paper tackles the methodologically and theoretically far-reaching consequences of the calculation of the h-point in lemmatised and non-lemmatised texts. After a short theoretical discussion and deductive derivation of the behaviour of the h-point in lemmatised and non-lemmatised texts, some empirical results from Russian, Slovene and Ukrainian will be discussed.

2. Lemmatisation – tokenisation: Quantitative and qualitative consequence

Word frequencies are studied in different branches of linguistics, in particular in quantitative stylistics, quantitative text analysis, psycholinguistics, authorship attribution and corpus linguistics. In word frequency studies, text pre-processing is obligatory and usually either word-form tokens or word-form lemmas are analysed. Whereas a study of word-form tokens is often based on a plain text approach, lemmatisation requires additional linguistic treatment of plain texts and the determination of standardised word forms, i.e. lemmas (e.g. infinitive or some basic form for verbs, nominative singular for nouns, etc.). Below, these two principally diverging approaches will be exemplified based on this Russian text:

> Kto iz vas, podlecov, kurit? Vse četvero ticho otvetili: - My ne kurim, batjuška. Lico popa pobagrovelo. - Ne kurite, merzavcy, a, machorku kto v testo nasypal? Ne kurite? A vot my sejčas posmotrim! Vyvernite karmany! Nu, živo! Čto ja vam govorju? Vyvoračivajte!
> [original Cyrillic text: Кто из вас, подлецов, курит? Все четверо тихо ответили: - Мы не курим, батюшка. Лицо попа побагровело. - Не курите, мерзавцы, а, махорку кто в тесто насыпал? Не курите? А вот мы сейчас посмотрим! Выверните карманы! Ну, живо! Что я вам говорю? Выворачивайте!]

Indeed text processing heavily depends on the problem of the word definition in general – a discussion which cannot be explored in detail here (for some recent discussions cf. Dixon, Aikhenvald 2002).

Quite often in quantitative and computational linguistics word definitions based on orthographic criteria are used. In this case usually the blank and punctuation marks are understood as a word delimiting sign, thus alphanumeric signs between two blanks in a written text are defined as one word form. Using this orthographic word definition for the Russian text presented above, the count yields the following results: 35 word-form types (counting word forms without taking their frequency into account) and 41 word-form tokens are obtained. In this case exactly five word forms appear more than once (*ne*, *a*, *kto*, *kurite*, *my*) and thus the number of word-form tokens is slightly higher than the number of word-form types. For details on the counted word form frequencies see Table 1.

Table 1
Frequency of word-form types.

No.	**Types**	**Frequency**	**No.**	**Types**	**Frequency**
1	*ne*	3	19	*kurit*	1
2	*a*	2	20	*lico*	1
3	*kto*	2	21	*machorku*	1
4	*kurite*	2	22	*merzavcy*	1
5	*my*	2	23	*nasypal*	1
6	*batjuška*	1	24	*nu*	1
7	*v*	1	25	*otvetil*	1
8	*vam*	1	26	*pobagrovelo*	1
9	*vas*	1	27	*podlecov*	1
10	*vot*	1	28	*popa*	1
11	*vse*	1	29	*posmotrim*	1
12	*vyvernite*	1	30	*sejčas*	1
13	*vyvoračivajte*	1	31	*testo*	1
14	*govorju*	1	32	*ticho*	1
15	*živo*	1	33	*četvero*	1
16	*iz*	1	34	*čto*	1
17	*karmany*	1	35	*ja*	1
18	*kurim*	1			

The lemmatisation (i.e. the reduction of word-form types to standardised word forms) leads again to a slightly lower number of counted lexical entities in the text. According to broadly accepted definitions (cf. Bußmann 2008: 296) the lemmatisation is the grouping together of different inflected forms to a standardised "basic form". Morphologically similar word forms are therefore analysed as one single item. During the lemmatisation, a disambiguation of word forms is

usually performed, too, and so homographs are identified and are treated in a linguistically more sophisticated way.

The quantitative consequence of a (manually performed) lemmatisation of the above Russian text is a shift of the frequency distribution of the counted lexical elements: the pronouns *vam*, *vas* are grouped together with *vy,* the inflected verb forms of *kurim* (1.P.Pl.), *kurit* (3.P.Sg.), *kurite* (2.P.Pl.) are reduced to the infinitive form *kurit'*. Finally, for the analysed text one obtains 32 lemmas, whereas in the case of tokenisation 35 word form types and 41 word tokens were obtained. For details on the counted lemma frequencies see Table 2.

Table 2
Frequency of lemmas.

No.	**Types**	**Frequency**	**No.**	**Types**	**Frequency**
1	*ne*	3	17	*lico*	1
2	*a*	2	18	*machorka*	1
3	*kto*	2	19	*merzavec*	1
4	*kurit'*	4	20	*nasypat'*	1
5	*my*	2	21	*nu*	1
6	*batjuška*	1	22	*otvetit*	1
7	*v*	1	23	*pobagrovet'*	1
8	*vy*	2	24	*podlec*	1
9	*vot*	1	25	*pop*	1
10	*vse*	1	26	*posmotret'*	1
11	*vyvernut'*	1	27	*sejčas*	1
12	*vyvoračivat'*	1	28	*testo*	1
13	*govorit'*	1	29	*ticho*	1
14	*živ*	1	30	*četvero*	1
15	*iz*	1	31	*čto*	1
16	*karman*	1	32	*ja*	1

In total, the following quantitative differences between lemmatisation and a simple tokenisation are obtained for the analysed Russian text: 41 word form tokens, 35 word form types and 32 lemmas (taking into account the frequency of lemmas equals the number of word-form types). From a purely quantitative point of view, simple tokenisation and lemmatisation are different approaches leading to a quantitative reduction of the lexical items in the texts. Beyond these quantitative consequences from a linguistic point of view, lemmatisation and tokenisation are rather different approaches, with a strong influence for any further qualitative interpretation of word form and lemma frequencies. Particularly one has to take into account that:

- Lemmatisation is predominantly relevant for inflecting languages. In this respect, the obtained quantitative differences between lemmatised and

non-lemmatised texts provide in-depth information about the inflectional activity of analysed languages. As a tendency, mainly nouns, verbs, adjectives, pronouns and adverbs are characterised by inflectional forms and thus these morphosyntactic word-form classes are mainly affected by lemmatisation.

- Lemmatisation provides a significant possibility to obtain the lexical richness of texts. The problem of lexical richness – a problem mainly discussed in quantitative stylistics, authorship attribution and quantitative text analysis – cannot be analysed in a satisfactory way without the lemmatisation of texts, especially in the case of languages with inflection.
- Disambiguation, which is usually performed during the process of lemmatisation, is a necessary precondition of a sophisticated analysis of the lexico-semantic structure of texts. Further, the disambiguation of texts gives detailed information about the number of homographs in a text – without lemmatisation[1] this information is not available to the linguist.

Generally, it is important to note that the question of tokenisation and lemmatisation of texts is directly dependent on the linguistic hypothesis being explored, although for stylistic purposes and related problems, one has to favour a lemmatisation of texts. In the next section the impact of lemmatisation on the determination of the h-point will be discussed in detail.

3. h-point: lemmatisation and tokenisation

The h-point, originally introduced into scientometrics and bibliometrics by Hirsch (2005), has recently been discussed intensively in quantitative linguistics and in word frequency studies (cf. Popescu 2007; Popescu, Altmann 2006; Popescu, Altmann 2007; Mačutek et al. 2007; Popescu, Altmann 2008). The h-point is a fixed point on a rank-frequency distribution, where the rank r and the frequency $f(r)$ of a countable linguistic entity coincide. For special cases, where one cannot obtain the point where $r = f_r$, one can determine the h-point by the point where the product of rank and frequency reaches its maximum (cf. Martináková et al. 2008: 93). In both the above-mentioned cases, the h-point can be obtained rather easily and mechanically. For the exact calculation of the h-point one can use the formula proposed by Popescu and Altmann (2008: 95):

$$h = \begin{cases} r & \text{if there is an } r = f_r \\ \dfrac{f_1 r_2 - f_2 r_1}{r_2 - r_1 + f_1 - f_2} & \text{if there is no } r = f_r \end{cases}$$

[1] Nevertheless, in some languages general problems of lemmatisation arise that must be solved by arbitrary decisions, e.g. in Hungarian, the verb lemma is given as the third person singular; or in Indonesian languages where there is no inflection, there are "basic words" from which everything else is derived.

where f_r is the frequency of the element at rank r.

According to Popescu et al. (2009), the h-point separates the vocabulary (V) of a text into two parts, namely into a class of magnitude h of frequent synsemantic auxiliaries (prepositions, conjunctions, pronouns, articles, particles, etc.) and a much greater class ($V - h$) of autosemantics, which are not so frequent but which build the very vocabulary of the text. Thus the h-point separates the "rapid" branch of synsemantics before the h-point from the "slow" branch of autosemantics after the h-point. Without a doubt the separation of autosemantic and synsemantic word forms is not clear-cut; sometimes there are autosemantics in the rapid branch and in other cases there are synsemantics in the slow branch. Thus the h-point is not an exact border between autosemantic and synsemantic word forms, but rather a fuzzy separating point of the lexico-semantic material on a rank frequency distribution.

The main field of application of the h-point is in quantitative text analysis and language typology. In cross-linguistic analysis and language typology the h-point can be interpreted as a sign of analytism, i.e. in analytic languages the number of word forms is smaller, and the synthetic elements are replaced by synsemantics. Furthermore, the h-point is considered to be a characteristic of individual texts within a given language and a sign of analytism/synthetism in cross-linguistic comparison. The h-point furthermore can be used for the measurement of the lexical richness of texts. This seems to be valid only when one accepts the area below the h-point (as the relevant one for the lexical richness of a text. As shown above, this area is characterised by a large number of autosemantics and thus Popescu et al. (2009: 95ff.) utilise this behaviour for their concept of the thematic concentration of texts.

Despite the broad applicability of the h-point, one has to take into account that the h-point systematically depends on text lengths. Therefore, one has to analyse texts which are approximately of similar length or one uses indices which are based on the h-point but normalised by text length (cf. Popescu et al. 2009: 19).

In any case the h-point plays a crucial role for word frequency studies and related problems of quantitative text analysis and cross-linguistic studies. All in all, the h-point seems to be an appropriate and multi-sided tool for word frequency studies, in particular text analysis and language typology. Whereas without any doubt, the h-point has a high conceptual value for the linguistic analysis of rank frequency distribution, to the best of our knowledge the influence of the lemmatisation of texts for the determination of the h-point has not been systematically analysed yet.

Deductively one can state that in the case of lemmatisation in highly inflectional languages the h-point will be higher than in cases of simple tokenisation, e.g. without lemmatisation. This is explainable by the fact – as already stated – that in the area before the h-point there are mostly synsemantics, which are as generally characterised by a low degree of inflection. In particular this holds true for synsemantics such as prepositions, conjunctions, particles, etc.,

which usually do not have inflectional forms. Furthermore, in the area before the h-point, the occasional autosemantic word forms (as types) are characterised by an extraordinarily high frequency (as tokens). A lemmatisation thus increases the frequency of these particular word forms because word forms of verbs, nouns, pronouns, etc. are reduced to a single item, e.g. one lemma. This behaviour has already been demonstrated in Section 1, where the frequency of particular word-form tokens like *kurit'* or *vy* increased. In the case of lemmatisation, the increase of the frequency of words forms causes a systematic shift of the h-point in the rank–frequency distribution. Or in other words: due to the lemmatisation and this systematic frequency shift of word forms to lemmas, the h-point "glides" to the right side on the rank frequency curve. In this case the lemmatisation of text thus causes an increase of the h-point, which should be rather systematic and in a systematic relationship to the length of the analysed texts. The empirical verification of these deductively found assumptions regarding the behaviour of the h-point in lemmatised and non-lemmatised texts will be performed on the basis of Russian, Slovene and Ukrainian texts in the next section. The focus is on a potential systematic shift of the h-point in lemmatised and non-lemmatised texts, using texts from different languages and of varying length.

3.1. Empirical evidence from Slovene

For Slovene the ranks and frequencies of word forms, i.e. word form tokens and lemmas, were determined in the short novel *Hlapec Jernej*, written by Ivan Cankar. The novel consists of ten chapters. All texts were first automatically lemmatised[2] by special software and in the second step all entries were checked manually. To get a more in-depth insight into the behaviour of the h-point in texts of different text lengths, the ten chapters were analysed in cumulative form (chapter 1, chapter 1 + 2, …, chapter 1 + 2 + … + 10).

Fig. 1a represents the behaviour of the h-point in cumulative Slovene non-lemmatised texts and in Fig. 1b the h-point in lemmatised Slovene texts can be seen.

First of all, the systematic behaviour of the h-point in relation to text length (= number of word forms and lemmas) can be obtained. This characteristic of the h-point is well known and explored in detail in Popescu et al. (2009: 19). Furthermore, it is obvious that the h-point in the lemmatised text is slightly higher than in the non-lemmatised texts (cf. Table 3 for details), but obviously due to the relative shortness of the analysed texts in some cases (chapters 1–9) no changes of the h-point can be seen, whereas in chapters 1–4 the h-point in the lemmatised texts is even lower than in the non-lemmatised texts.

[2] The software is available for free at http://nl.ijs.si/analyse/. The error rate for the analysed text is 5%.

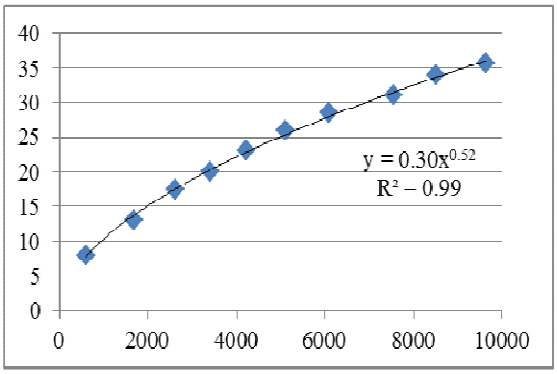

Figure 1a. The h-point in Slovene non-lemmatised texts.

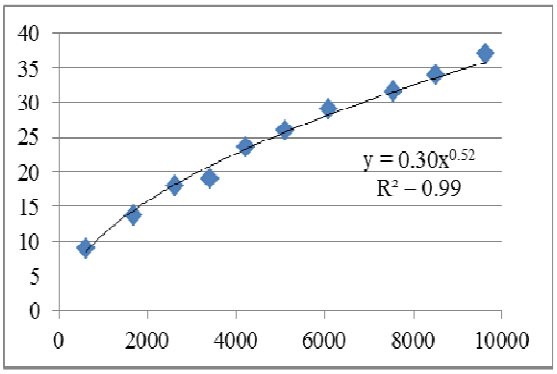

Figure 1b. The h-point in Slovene lemmatised texts.

Table 3
The h-point in lemmatised and non-lemmatised texts (Slovene).

Chapter	Tokens	Types	Lemmas	h-point	
				Tokens	Lemmas
1	602	343	283	8	9
1–2	1679	676	519	13	13.7
1–3	2617	938	687	17.5	18
1–4	3413	1149	814	20	19
1–5	4222	1302	910	23	23.5
1–6	5112	1477	1013	26	26
1–7	6085	1688	1133	28.5	29
1–8	7558	1969	1297	31	31.5
1–9	8497	2173	1404	34	34
1–10	9631	2361	1500	35.7	37

Thus we can state the following intermediate result for the Slovene text: as a rule of thumb, the h-point in lemmatised and non-lemmatised texts is in approximately the same position, or in other words, in our Slovene database no substantial differences can be obtained. This observation, which is in contradiction to our suggested behaviour of the h-point, confirms the already known systematic interrelation of the h-point with text length but not the systematic shift of the h-point. Nevertheless this finding seems to be explainable by the relative shortness of the texts used, which does not exceed 10,000 word forms and 1500 lemmas per chapter. Regarding the relevance of lemmatisation in quantitative text analysis and language typology, this can be understood in the following way: for a suitable interpretation at least some minimal text length is required and in the case of relatively short texts the lemmatisation does not provide any substantial additional linguistic information in comparison to a simple tokenisation.

3.2. Empirical evidence from Russian

Russian is usually considered to be a synthetic and strongly inflected language. The analysed Russian texts are slightly longer (cf. Table 4 for details) than the Slovene texts, so the impact of text length can be analysed in more detail.

The texts are taken from the Russian novel *Kak zakaljalas' stal'* (*How the Steel Was Tempered*). The lemmatisation of the ten analysed chapters was performed by TreeTagger, a lemmatiser for Russian available for free[3]. The novel *Kak zakaljalas' stal'* represents a specific form of the socialistic realism literary language of the 1930s and thus the lemmatiser we used does not recognise many of the word-form tokens in a proper way. The success rate was 90 per cent and therefore over 4,000 types were lemmatised manually.

The results of determining the h-point in lemmatised and non-lemmatised texts (cf. Table 4) show that the h-point in lemmatised texts is – as already predicted – in all cases higher than in non-lemmatised texts. Again, the predicted systematic relationship between the h-point and text length is found.

Table 4
The h-point: Non-lemmatised and lemmatised texts (Russian).

Chapter	Tokens	Types	Lemmas	h-point	
				Tokens	Lemmas
1	4107	1907	1376	20	21.7
1–2	8243	3538	2407	25.7	31
1–3	14567	5591	3605	35	41
1–4	18300	7003	4435	38.4	45
1–5	22070	7956	4931	42	51
1–6	28709	9822	5878	50	57
1–7	33940	11388	6693	55	63
1–8	40979	12864	7435	59	68
1–9	44275	13635	7814	62	71
1–10	49678	15053	8495	63	76.5

A graphical representation of the systematic behaviour of the h-point in lemmatised and non-lemmatised texts can be found in Fig. 2.

[3] I would like to thank Ruprecht von Waldenfels (University of Bern) for his help with the automatic lemmatisation of the Russian texts.

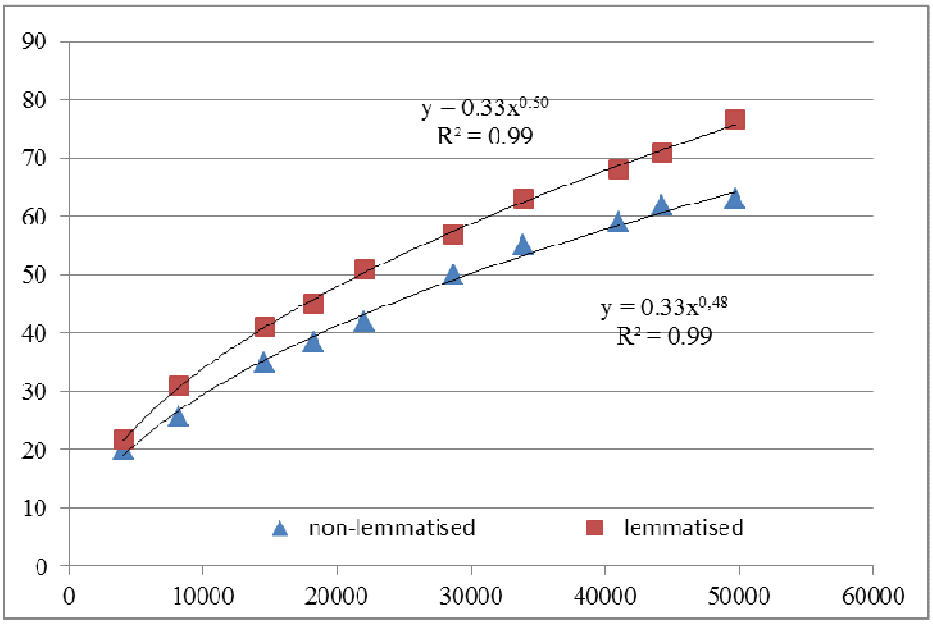

Figure 2. Text length vs. h-point.

As can be seen from Fig. 2, the distance between the h-point in lemmatised and non-lemmatised texts increases quite systematically with text length: The longer the texts, the larger the distance between the h-point in lemmatised and non-lemmatised texts. Generally, the analysis of the Russian texts confirms the deductively found claims about the h-point in lemmatised and non-lemmatised texts and furthermore the need for an analysis of texts with a suitable length is confirmed.

3.3. Empirical evidence from Ukrainian

Finally, the results for the Ukrainian texts can be presented. Compared to the Russian and Slovenian data, no cumulative chapters but eight individual novels written by the Ukrainian writer Ivan Franko were analysed. The corpus consists of the following titles (Buk 2007; 2013): *Boa constrictor* (1st edition: 1878–84; 2nd edition: 1905–07), *Boryslav smijetsja* (Boryslav Laughs) (1880–81); *Zakhar Berkut* (1883); *Ne spytavšy brodu* (Without Asking a Wade) (1885–86); *Dlja domašnjoho ohnyšča* (For the Hearth) (1892); *Osnovy suspil'nosty* (Pillars of Society) (1894–95); *Perekhresni stežky* (The Cross-paths) (1900); *Velykyj šum* (The Great Noise) (1907); *Petriji j Dovbuščuky* (2nd edition: 1909–12). In this work, the titles are referred to by the first letters of the Ukrainian transliteration, i.e.: BC, BC2, BS, ZB, NSB, DDO, OS, PS, VS and PD2. The texts of the corpus are tagged with part-of-speech and lemma information provided for each token. The tagging was performed semi-automatically. First, manual disambiguation of homographs was made. In the second step, the dictionary of word-form types

was created and for each type the part-of-speech and lemma were given. This dictionary was used to tag every subsequent text and expanded as necessary.

The word forms were lemmatised, i.e., reduced to an initial (vocabulary) form: verbs to the infinitive, nouns and pronouns to nominative singular, adjectives to masculine nominative singular, etc. Suppletive forms of adjectives and adverbs (superlative and comparative) were reduced to separate (comparative) forms. Another point to be noted is the euphonic changes, which are also taken into consideration (cf. Buk, Rovenchak 2007). This affects mostly *і/ǔ* and *в/у* alternation (word-initially and as separate words) and the verbal reflexive particle *-ся/сь*. As a rule of thumb, the vowel variant (*і* [i], *у* [u]) is used between consonants, and the consonant variant (*ǔ* [j], *в* [v]) is used between vowels. In a mixed phonetic environment the choice is less strict as it is conditioned by phonetic harmony. The respective lemmas were joined into one type in frequency lists of all the texts apart from BC2 and PD2. The latter two novels were not included in the comparison but the preliminary analysis suggests that no substantial difference is observed: a slight shift of the h-point is caused by the fact that *і/ǔ* and *в/у* lemmas have rather high frequencies with ranks $r < 10$. All data for the Ukrainian texts can be found in Table 5.

Table 5
The h-point in Ukrainian texts.

Texts	**Tokens**	**Types**	**Lemmas**	**h-Point**	
				Tokens	**Lemmas**
BC	25427	8351	5007	48	56
BS	77456	16069	8572	98.5	109.5
DDO	44841	11518	6472	71.5	78
NSB	49170	12808	7140	73.8	80
OS	67173	15437	8345	89.3	101.5
PS	93884	19425	9961	105.5	113.5
VS	37005	11058	6468	62	67
ZB	50206	12494	6520	75.8	87

It can be seen that the h-point in the Ukrainian texts (cf. Fig. 3) shows the same behaviour as already obtained in the Russian texts: (a) the h-point in lemmatised texts is slightly higher than in non-lemmatised texts and (b) in both cases the h-point systematically interrelates with the text length.

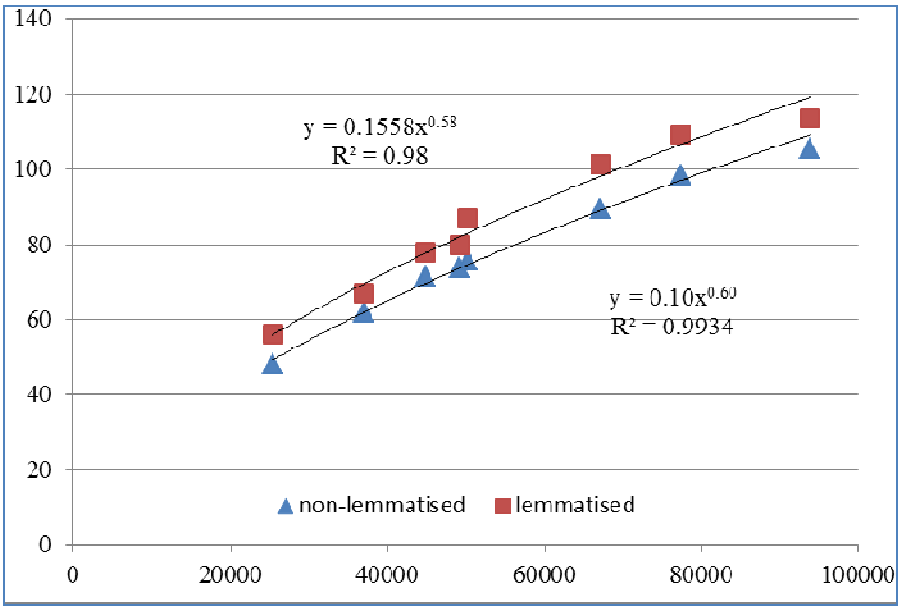

Figure 3. The h-point in Ukrainian novels

4. Summary

The main results of the systematic comparison of the h-point in lemmatised and non-lemmatised texts are as follows:

(1) In two analysed languages (Russian, Ukrainian) the h-point shows the deductively predicted behaviour. The h-point systematically interrelates with the text length, regardless of the counted lexical units (word-form tokens or lemmas).

(2) In the Russian and Ukrainian lemmatised texts without any exception, the h-point is higher than in the non-lemmatised texts, so one can claim a rather systematic effect of lemmatisation on the h-point.

(3) For Slovene no clear-cut results can be obtained, i.e. the h-point is equal in lemmatised and non-lemmatised texts. Considering the results of the Russian and Ukrainian texts, this quite specific behaviour can be explained by the relatively short length of the analysed texts. Hence in addition to the kind of text processing (lemmatisation or simple tokenisation) in h-point studies, one has to consider the sample size of analysed texts.

Altogether, the results obtained from Russian, Ukrainian and Slovene texts support the general relevance of the h-point for word frequency studies. In addition to its central function of being a fuzzy border between autosemantic and synsemantic word forms of rank–frequency distribution, which are based on word-form tokens, the h-point can indeed serve as a useful indicator for the lexical richness of texts. The latter holds true especially when sufficiently long texts

are used (such as the described Russian and Ukrainian texts), whereas in relatively short texts (e.g. the Slovene database) no substantial differences of the h-point in lemmatised and non-lemmatised texts could be obtained. Thus in short texts – at least the results from the Slovene texts support this view – the h-point can be understood as a general tool for cross-linguistic comparison, stylistic analysis and particularly lexical richness as the lemmatisation of texts does not affect the position of the h-point substantially.

References

Bußmann, H. (2008). *Lexikon der Sprachwissenschaft. Vierte, duchgesehene und bibliographisch ergänzte Auflage.* Stuttgart: Kröner.

Buk, S. (2007). Korpus tekstiv Ivana Franka: sproba vyznačennja osnovnykh parametriv [Ivan Franko text corpus: an attempt to define main parameters]. In: Širokov, V. A. (ed.), *Prykladna linhvistyka ta linhvistyčni tekhnolohiji: MegaLing-2006: 72–82*. Kyiv: Dovira.

Buk, S. (2013). Kvantytatyvna parametryzacija tekstiv Ivana Franka: proekt ta joho realizacija [Quantitative parametrisation of texts written by Ivan Franko: the project and its realization]. *Visnyk Lvivs'koho universytetu. Serija filolohična 58, 290–307*; see also preprint arXiv:1005.5466v1 [cs.CL].

Buk, S., Rovenchak, A. (2007). Statistical parameters of Ivan Franko's novel Perekhresni stežky (The Cross-Paths). In: Grzybek, P., Köhler, R. (eds.), *Exact Methods in the Study of Language and Text. Dedicated to Gabriel Altmann on the Occasion of his 75th Birthday: 39–48.* Berlin, New York: de Gruyter.

Dixon, R.M.W., Aikhenvald, A.Y. (2002). Word: a typological framework. In: Dixon, R.M.W., Aikhenvald, A.Y. (eds.), *Word. A cross linguistic typology: 1–41*. Cambridge: Cambridge University Press.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the USA 102(46), 16569-16572*.

Mačutek, J., Popescu, I.-I., Altmann, G. (2007). Confidence intervals and tests for the h-point and related text characteristics. *Glottometrics 15*, *45–52*.

Martináková, Z., Mačutek, J., Popescu, I.-I., Altmann, G. (2008). Some problems of musical texts. *Glottometrics 16*, *80–110*.

Popescu, I.-I. et al. (2009). *Word Frequency Studies.* Berlin, New York: Mouton de Gruyter.

Popescu, I.-I. (2007). The ranking by the weight of highly frequent words. In: Grzybek, P., Köhler, R. (eds.), *Exact Methods in the Study of Language and Text. Dedicated to Gabriel Altmann on the Occasion of his 75th Birthday: 555–565.* Berlin, New York: de Gruyter (Quantitative Linguistics, 62).

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

Popescu, I.-I., Altmann, G. (2007). Writer's view of text generation. *Glottometrics 15*, *71–81.*

Popescu, I.-I., Altmann, G. (2008). On the regularity of diversification in language. *Glottometrics 17*, *94–108.*

Empirical Approaches to Text and Language Analysis

dedicated to Luděk Hřebíček on the occasion of his 80th birthday

edited by

Gabriel Altmann, Radek Čech,
Ján Mačutek, Ludmila Uhlířová

2104
RAM-Verlag

Studies in quantitative linguistics

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

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

ISBN: 978-3-942303-24-8

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

Contents

Addresses of authors

Altmann Gabriel
Lüdenscheid, Germany
ram-verlag@t-online.de

Andreev Sergey
Smolensk State University, Smolensk, Russia
smol.an@mail.ru

Andres Jan
Palacký University, Olomouc, Czech Republic
jan.andres@upol.cz

Buk Solomija
Ivan Franko National University of Lviv, Lviv, Ukraine
solomija@gmail.com

Čech Radek
University of Ostrava, Ostrava, Czech Republic
cechradek@gmail.com

Fan Fengxiang
Dalian Maritime University, Dalian, China
fanfengxiang@yahoo.com

Huang Wei
Beijing Language and Culture University, Beijing, China
huangwei@blcu.edu.cn

Kelih Emmerich
University of Vienna, Vienna, Austria
emmerich.kelih@univie.ac.at

Köhler Reinhard
Universität Trier, Trier, Germany
koehler@uni-trier.de

Kubát Miroslav
Palacký University, Olomouc, Czech Republic
miroslav.kubat@gmail.com

Liu Haitao
Zhejiang University, Hangzhou, China
lhtzju@gmail.com

Mačutek Ján
Palacký University, Olomouc, Czech Republic
jmacutek@yahoo.com

Mácha Jiří
Charles University, Prague, Czech Republic
jiri.macha.84@gmail.com

Melka Tomi S.
Parkland College, Champaign (IL), USA
tmelka@gmail.com

Mikros George K.
National and Kapodistrian University of Athens, Athens, Greece
gmikros@gmail.com

Milička Jiří
Palacký University, Olomouc, Czech Republic
Charles University, Prague, Czech Republic
milicka@centrum.cz

Richterová Olga
Charles University, Prague, Czech Republic
richterova.olga@gmail.com

Rovenchak Andrij
Ivan Franko National University of Lviv, Lviv, Ukraine
andrij@ktf.franko.lviv.ua

Sanada, Haruko
Saitama Gakuen University, Saitama, Japan
h_sanada@nifty.com

Stachowski Kamil
Jagiellonian University, Cracow, Poland
kamil.stachowski@gmail.com

Su Hong
Dalian Maritime University, Dalian, China
suzihesuhong@126.com

Uhlířová Ludmila
Czech Academy of Sciences, Prague
lidauhlirova@seznam.cz
uhlirova@ujc.cas.cz

Wang Lu
Universität Trier, Trier
wanglu-chn@hotmail.com

Wimmer Gejza
Matej Bel University, Banská Bystrica
Slovak Academy of Sciences, Bratislava
wimmer@mat.savba.sk

Zhou Pianpian
Dalian Maritime University, Dalian
zhoubianer@163.com