SEASONALITIES IN THE INTRODUCTION OF WORD-TYPES IN LITERARY WORKS

MARIA CSERNOCH

1. Introduction

Since the fully automated processing of texts written in any natural language is not solved, one of the possible approaches to the problems is to reduce their complexity by selecting a feature of the text and giving a detailed description of this phenomenon. One of these reductions is to use the obvious simplification that words occur randomly, more precisely, independently in texts. Until now several promising results have come to life which all used these simplifications. Researches were mainly focusing on vocabulary size and richness and were trying to find formulae which are able to give reliable pieces of information about these characteristics of the texts.

Along with the randomness assumption another simplification had to be applied, namely that words in a text belong to the Large Number of Rare Events (LNRE) zone. Until now, however, no really good and fast algorithm has been found to model these distributions, so it is usually assumed that words are multinomially, or, as a special case, binomially distributed.

Applying one of these models Baayen (Baayen 1996a, Baayen 1996b, Baayen 2001) came to the conclusion that the randomness assumption is violated not on sentence, but either on paragraph or discourse level. He also had the suggestion that the constraints on discourse level might be responsible for the differences between the original and the expected vocabulary size, and also gave a vague explanation.

The primary aim of this study was to analyze a previously not mentioned parameter, the appearance of word-types in literary works. This was carried out using some of the previously applied theories and methods of computer-aided lexical statistical analyses and by introducing some novel methods to fit to this special problem. We mainly worked with novels and concatenated short stories in English and Hungarian, searching for explanations, reasons why, when, how many etc. relatively new words are introduced into the texts.

It was known from previously published studies that the analysis of word-types on its own is not perfunctory for identifying the authors. Aware of this fact we were looking for parameters which can be gained from the appearance, not from the number of the word-types.

Most of us would agree that reading a book becomes easier and easier as the story goes on, as we are heading towards its end. This feeling in case of our mother tongue is not as clear as with texts written in a foreign language. All this was

proved by counting the words and following the changes in vocabulary size. It also became evident after some polls that although a single measure of vocabulary richness, that can characterize an author or a text, is an attractive idea, however, reader's perceptions about vocabulary richness are not necessarily accurate (Hoover 2003).

The number of the newly introduced word-types in a text shows, in general, a monotonic decay. On the other hand, in most of the texts we can find intervals, parts of the texts, where this monotonic decay is reversed and a sudden increase in the number of the newly appearing words can be detected (Csernoch 2004; Csernoch 2006).

In our study we wanted to give explanations for these sudden increases, aiming to find reasons why authors use in these text-slices more words than they did previously. We also wished to see whether these changes are predictable, if there is any regularity in their appearance or not.

To carry out our experiments we had to build a dynamic model which is able to give a good approximation of the original text in its progress. The other constraint on the model was that it should be language free, it should be able to work with texts written in different languages. Our main goal was to analyze both English and Hungarian texts. English texts were chosen to obtain results that are directly comparable to previously published works, Hungarians to see how an agglutinating language can be modelled and get comparable, if there are any, results to texts in English and in any other languages. The emphasis was put on the dynamic characteristic of the model, which should be able to reproduce at least the trends but most preferably also the seasonalities of the original texts.

Previously published works, mainly literary criticism, showed clearly that even experts of the field do not share their opinions on when, at which place of the text new words appear. This is to some extent understandable since their opinions are mostly based on impressions. Some of them thought that the boundary of the chapters is where sudden increase in the number of the word-types can be detected (Balázs 1985). On the other hand, in other's opinion a change occurs when there is an interruption in the flow of the text, i.e. where a text-slice is inserted which differs in style from the text as a whole, e.g. a longish description appears unexpectedly (Genette 1980).

Familiar with these opinions and also with Baayen's results and expectations (Baayen 2001) we wanted to prove the hypothesis that the number of newly introduced word-types increases when a sudden change can be detected in the flow of the text. These changes are relatively short, compared to the length of the whole text, but clearly separable. This statement can be rephrased: the differences between the original and the artificial texts, created by the dynamic model, are due to changes at discourse level of the original text, neither the changes on sentence nor paragraph level cause measurable changes in the number of word-types.

In our works beyond building the dynamic model we also applied a method hardly ever used in lexical statistical studies, the comparison of the original text and its translations. This method did not seem applicable in earlier works because they focused on the overall number of words, and texts with different vocabulary size cannot be compared, at least not easily. The difference in vocabulary size is due to both the characteristics of the languages and the translator's freedom. Since our aim was not to count or give prediction for the number of the words but to follow the changes of the words in progress, the previously considered problems did not cause difficulties or give any obstacles to carry out our experiments.

2. Literary works mentioned in this study

2.1. Works in their original languages

Bierce, Ambrose: *An Heiress from Redhorse, The Man and the Snake*; Conrad, Joseph: *Youth, The End of the Tether,* and *Lord Jim*; Crawford, Marion F.: *By the Waters of Paradise*; Dickens, Charles: *Great Expectations, American Mystery Stories*; Freeman, Mary E. Wilkins: *The Shadows on the Wall*; Hawthorne, Nathaniel: *The Scarlet Letter*; Irving, Washington: *Wolfert Webber or Golden Dreams*, *Adventure of the Black Fisherman*; Kertész Imre: *Sorstalanság* (for which Imre Kertész was awarded the Nobel Prize in Literature in 2002); Kipling, Rudyard: *The Jungle Book*; London, Jack: *An Odyssey of the North, Call of the Wild*, and *Seawolf*; Poe, Edgar Allan: *The Oblong Box, The Gold-Bug*; Post, Melville D.: *The Corpus Delicti*; Rowling, J. K.: *Harry Potter and the Philosopher's Stone*; Twain, Mark: *The Adventures of Tom Sawyer* (from now on: *Tom Sawyer*).

2.2 Translations

The German: *Roman eines Schicksallosen*[1] and English: *Fateless*[2] translations of Imre Kertész's *Sorstalanság*.

3. Methods

3.1. Data retrieval from texts

Our main goal was to analyze novels and short stories written in English and Hungarian. For the analysis we needed the electronic versions of the original, printed texts. The main source of these electronic versions was the Internet. The

[1] Aus dem Ungarischen von Christina Viragh (1996), Rowohlt Taschenbuch Verlag, Hamburg

[2] Translated by Christopher C. Wilson and Katharina M. Wilson (1992), Hydra Books, Northwestern University Press, Evanston/Illinois

texts that were not available free through the Internet were scanned manually. It should be noted here that the availability of electronic versions greatly influenced the selection of works that were finally included in the present study.

To carry out the experiments a software, *DyMoCASAT* (Dynamic Models for Computer Aided Statistical Analysis of Texts) was developed. *DyMoCASAT* carries out the data retrieval from the original text, the building of the model, and based on the model the generation of the corresponding artificial texts.

DyMoCASAT has two character sets by default: English and Hungarian. Any other character sets can be set up within the program offering access to texts written in other languages.

Since our final goal was to gather information about the appearance and the behaviour of the word-types in literary texts, the starting point of our experiments had to be the definition of words (word-types). First, the character set, the alphabet, was determined upon which the program is able to decide which string is a word (type) and, based on this crucial information, were all the experiments carried out. Since pre-processing was not applied to any of the texts, the word-type, a string of characters between two separator characters, was declared as the basic unit of the analysis.

3.2. Storing data

The analysis of the texts had to be preceded by saving all the available information about the number and the exact places of the word-types. This was all carried out automatically by *DyMoCASAT*. In contrast to previously published works, we were to examine the appearance of the word-types in progress. Since the number of the newly introduced word-types is greatly influenced by the length of the intervals in question, intervals of different lengths could not be used. In this respect our model differs from those presented earlier, where texts are divided into equal number of intervals independent of the length of the given text, giving rise to text-slices of different lengths. Here, instead, we kept the lengths of the intervals and blocks constant (h). As a result the number of the blocks changes from text to text. To use this novel approach a suitable constant for the length of the intervals had to be chosen.

Usually blocks containing one-hundred tokens ($h = 100$) were chosen. Two advantages of these short blocks of constant-length were found over the previously used method. First, since the length of a block is independent of the length of the original text, the individual slices from different texts can be compared readily (Table 1).

The second advantage of using hundred-token-long blocks comes from the relatively short length of these blocks. Using these short blocks subtle changes, couple of hundred-token-long text-slices, in the narrative can also be traced (Fig. 1).

The following variables are used in *DYMOCASAT* and, consequently, in this study:

N	the number of tokens in a text, the sample size,
$V(N)$	the size of the vocabulary in an N-token-long text, the number of the different word-types,
ω_i	the i^{th} word in a list of word-types ordered by frequency,
$f(i,N)$	the frequency of ω_i in a sample size of N token,
h	the length of the intervals (blocks) into which the text is divided,
b_i	the i^{th} block, by dividing the text into h-token-long blocks,
y_i	the number of the newly introduced word-types in the i^{th} block: $f(b_i) = y_i$ (Fig. 2C and D)
n	the number of the blocks, using h-token-long blocks.
$V(b_i)$	the cumulative vocabulary size, when i reaches n $V(b_n) \cong V(N)$ (Fig. 2A and B)

$$b_i,\ i = 1, \quad , n, \text{ where } n = \left\lceil \frac{N}{h} \right\rceil \tag{1}$$

The method of dividing the texts into h-token-long slices always produces some loss, since the text at the end is truncated to n h words.

The loss is minor, $v = N - h \cdot \left\lceil \frac{N}{h} \right\rceil$, compared to the size of the texts, so it will not influence the results of our experiments.

3.3. Building the model

Models based on the frequency of words assume that the words appear randomly within texts. There are, however, a number of strategies how random selections can be carried out (for review see Baayen 2001). The best results were obtained with models that assume that word-types follow the multinomial distribution, since multinomial distribution arises when each trial has k possible outcome. Selecting word-types from a set of tokens is exactly the same problem, where the number of the possible outcome is $V(N)$, the number of the different word-types in an N token long text.

Our model also uses the frequencies of the word-types ($f(i,N)$) of the original text, and their relative frequencies

$$frel(i,N) = \frac{f(i,N)}{N} \in \left]0;1\right[, \tag{2}$$

thus the probability of occurrences (p_i). While previous works focused on the overall vocabulary size ($V(N)$) and richness, the given formulae were able to produce reliable pieces of information (for review see Baayen 2001). However, our

aim was not the determination of the vocabulary size, rather to find trends or trace seasonalities, if there are any, in the text flow. The previously given formulae are not able to provide information about a text in progress. Given these constraints new methods with new theoretical background had to be found (Csernoch 2006).

3.3.1. Selecting the words without replacement (H)

For this model the tokens of the texts were stored in a one dimensional array. The tokens were randomly picked from this array, but after checking and saving their types they were not put back.

3.3.2. Trends in the appearance of the word-types

After counting and storing all the occurrences of the words, the program plots the number of newly introduced word-types in each block ($f(b_i) = y_i$, $i = 1, n$) (Fig. 1, Fig. 2C and D, Fig. 3.). The number of the newly introduced word-types, in general, follows a decaying tendency. There are, however, parts in the texts where their number is greater than what is expected from the general trend. A point or a group of points that fall significantly outside of the general trend and form a local maximum within the neighbouring blocks will be referred to as a protuberance. As mentioned earlier, the protuberances on the graphs of the newly introduced word-types are visible only if h was defined appropriately.

It is clear that the number of the newly introduced word-types follows a generally decaying tendency (as mentioned by e.g. Muller 1964, Holmes 1994, Baayen 1996a, 1996b; for review see Baayen 2001) with an appreciable amount of noise. For detailed comparisons it was necessary to reduce this noise. To this end a 7-point smoothing with a second order polynomial (Scarborough 1966) and a Gaussian weighting function was used (*SIGMAPLOT*, SPSS Inc.).

Filtering the graph of the original text gave rise to a decaying function on which the smaller and larger secondary humps were now clearly visible. To decide which of these peaks stand for significant changes and which are due only to the noise of how the author selected the word-types the smoothed original graph (fp) and the average function of the artificial texts ($F(b_i) = Y_i$) were compared. The difference between the smoothed original and the average artificial text was determined and plotted (fp-F), and the mean (M) and the standard deviation (σ) of the difference function were calculated. Those differences are considered as significant which reach the $M + 2\sigma$ values.

4. Results

4.1. Literary works in English and Hungarian

The dynamic model created with the above detailed method was able to give account for the appearance of the word-types not only in English but also in Hungarian (and also in German) texts. It was found that the number of word-types is indeed higher in Hungarian texts due to the morphologically productive nature of the language (Table 2 and 3). The monotonic decay of the graphs of the newly introduced word-types and the noise on the graphs, however, follow the same pattern as found in English texts. Again, due to its productive nature the noise, as expected, was greater in the Hungarian texts. On the other hand, Hungarian language does not have fix word-order within sentences. As a result, theoretically any order of the words can be acceptable with a change in the focus (Table 4). Consequently, the randomness assumption should be more effective in describing the introduction of word-types in Hungarian (and German) texts as it was found for English texts. Texts of similar lengths regardless of the differences between the two languages did not show greater differences between the original and the artificial text than with the English texts.

This result meant that this analysis is language independent, thus for further investigation there is no need for different models for different languages. Texts written in different languages can be analyzed with the same method, greatly simplifying the comparison of these texts.

4.2. Comparison of the original and the artificial texts

The number of newly introduced word-types has, as expected, a general tendency to decrease along the course of the narrative. Beyond this decaying tendency the number of newly introduced types can, in many cases, be more at a later point in the discourse than in a previous section. These sudden changes cause smaller or greater protuberances on these graphs.

In the analyzed works not only the intensity but the length of the protuberances are also different, so shorter or longer rising phases were observed that interrupted the otherwise declining function of the newly introduced word-types. The subtle changes in the discourse were visible only if the intervals into which the texts were divided were short enough. The graphs, of course, show not only those subtle changes, marked by primary peaks, that were coming from the logical flow of the story but also those, where the text contains parts which are only slightly related to the events, causing secondary peaks. In both cases the monotonic decay of the graphs of the newly introduced word-types was somewhat reversed.

The question arises whether the randomness assumption (Mandelbrot 1962, Carroll 1967, Sichel 1986, Baayen 1996a) is able to account for these changes in

the course of graphs or not. As it was found, our dynamic model was able to simulate the feature marked by the primary peaks remarkably well. On the other hand, the model was incapable of reproducing the secondary peaks, which are, as mentioned earlier, totally unpredictable. The question was to give an explanation for the emergence of these secondary peaks. Fortunately, however, events for which the model was not able to give clear account can be traced back in the original text with the same model.

Relying on the readers' intuition the changes in the texts which produce the protuberances on the graphs of the newly introduced word-types should mainly coincide with the launch of a new chapter. Contrary to these subjective opinions Genette (1980) gives a detailed, but still subjective analysis of several literary works concerning the changes and the results of these changes in the flow of the texts. The above listed reasons provided, in most cases, more significant changes in the number of the newly introduced word-types than the introduction of a new chapter. His findings, concerning changes in lexis and grammatical markers, were supported by applying our method.

We found that the model used on literary works was able to follow the overall monotonic decay of the newly introduced word-types. The size of the noise on the graphs of the artificial texts was also similar to that of the original text, if only the general noise is considered. These changes which were due to the flow of the story gave rise only to small peaks in the otherwise decaying graph. However, the trace of those surprising, unrelated events, understandably, never occurred in the model. We were interested whether, by comparing the original text and the related model, we can pinpoint parts which are only loosely related to the story, the pure narrative of the story.

We selected texts of different genres, lengths, authors, languages to show that none of them influences the comparison of the original and their corresponding artificial texts. We have chosen for this comparison an English (*Tom Sawyer*), and a Hungarian novel (*Sorstalanság*), and two collections of short stories. One of them is *The Jungle Book* and the other is a collection of *American Mystery Stories* from different authors.

Analyzing the protuberances of *Tom Sawyer* we found two which hardly reached and one that exceeded the threshold of significance (Fig. 4A). The first stands for a prayer, the second is a daydreaming, while the third one is large and corresponds to a section where a school year examination is detailed, for which each student had to write a short story or a poem. The first of these little writings raised the number of newly introduced word-types significantly, which was kept high until all the little stories had been read. The style and vocabulary of these little pieces of writing are clearly different from the rest of the novel, not Tom-sawyerish at all, while the first two is not as far off from the style of the whole text.

In Imre Kertész's *Sorstalanság* six peaks were found which exceeded the

threshold of significancy (Fig. 4B). All of these protuberances occurred at places where longish descriptions were introduced into the text. Rereading these text-slices it was again found that neither of them carries vital information, information which is necessary to understand the story. All of them are used to give highly detailed description of newly introduced settings, situations, humans, etc., which barely have any straightforward connection with either the previous or the following events. These six, only slightly related events to the main course of the story of the text are the following (in order of appearance): the description of the first concentration camp, arrival to the second camp, start of the day and the work in the factory, the hospital, Pjetka's cooking, preparations for leaving the camp, including songs which are hyphenated at the syllables. Some of these events are major in the course of the story but the long descriptions which accompanied them use a high number of words which are unfamiliar with the style of the text. This use of vocabulary causes the protuberances on the graph and difference between the original and the artificial texts.

The *Jungle Book* contains seven tales and poems but only five peaks were identified on the difference trace (Fig. 4C). From these five three corresponded to the start of a new tale. These tales, White Seal, Rikki-Tikki-Tavi and Toomai of the Elephants, introduce new sceneries, especially evident for the White Seal, and thus give explanation for the sudden increase in the newly introduced word-types. It should be noted that neither the other four tales nor the songs appeared as peaks on the trace. The first peak, which is the most easily separable from the start of a tale, is the most characteristic. It is a long description of the Kings' Palace, the setting did not change, we are still in the jungle but there is a change in the register, which produces this huge increase in the number of word-types.

The concatenated *American Mystery Stories* gave three distinguishable peaks from which only one coincided with the start of a new story (Fig. 4D). All three correspond to a relatively long description. Similarly to the *Jungle Book*, it was found that the beginning of a new story does not necessarily give rise to the number of the word-types, only in case when the story starts with a longish description of a setting. What was, however, remarkable that even the change in author did not produce protuberances. The one peak which coincides with the beginning of a story marks a long description at the beginning of E. A. Poe's *The Gold-Bug*. The interesting feature of this peak is that the story previous to this one is also by Poe, which further strengthens our hypothesis that the genre overrides the author if the behavior of word-types is in question.

These observations clearly establish that the secondary peaks observed on the graphs of newly introduced word-types correspond to events, descriptions only loosely related to the discourse and, furthermore, even a text from a new author does not necessarily increases the number of newly introduced word-types.

It might be surprising that neither in *The Jungle Book* nor in *American*

Mystery Stories did the connecting points of the stories cause measurable peaks. From previously published works (for review see Baayen 2001) it is well known that the size of the vocabulary at any point of the text is not influenced by the overall length of the story. To demonstrate that not only the vocabulary size but the introduction of word-types is independent of either the overall length of the text or the author we picked nine stories of different lengths and authors (Fig. 5). Knowing all these and that the number of the newly introduced word-types are still high when the short stories are concatenated is not surprising any more that simply connecting them, the connection itself does not cause measurable differences in the number of the newly introduced word-types. Since there is no remarkable difference between how authors add new word-types to their stories even short stories from different authors do not produce protuberances at the connection points.

4.3. Comparison of the original text and its translations

To get comparable results we chose works whose translations were also available (at least in printed form). The original Hungarian work of Imre Kertész, its English, and German translations were analyzed. The question was again whether the difference between the original and the artificial texts are due to constraints on discourse level or below it, either on sentence or paragraph level.

If the difference between the original and the artificial text had been caused by syntactic or semantic constraints the translations would have presented protuberances on the graphs of the newly introduced word-types on totally arbitrary places. But this was not the case.

The data gained by analyzing the texts in different languages show that there are great differences between the number of the word-types, the number of the hapax legomena of the works and their translations, which are due to the differences between the languages and the translators' freedom. This method can produce comparable results only in cases when the differences are due to changes on discourse level.

In the German translation of *Sorstalanság* we found seven protuberances (Fig. 6). The first appeared earlier than in the Hungarian text and described the process of loading the people into the train not causing any significant change at arriving to the concentration camp, which is reasonable because for the two events similar vocabulary is applied. The second and the third protuberances occurred at same events in both texts. The fourth peak of the German text is the description of the protagonist's mental and physical status, which in the Hungarian text only caused a minor peak not reaching the threshold level. Finally, the last three appeared exactly on the same place as in the Hungarian text.

We got similar results analyzing the English translation, *Fateless*. Again, the protuberances were detected at those places where longish descriptions are inserted

into the text. Most of these places are identical to those found in the Hungarian and the German texts, but we have at this time eight peaks (Fig. 7). When compared them to the Hungarian and the German texts we found two at the very beginning which were not present in either of them, but they are descriptions. The next four are identical to their corresponding texts-slices in the other two texts, while two at the end did not appear in them. However, examining two small peaks in the English text that almost reach the threshold and are just prior to the last two we find that these are the same events as in the two other texts.

The question was why these slight differences appear in the English text compared to the other two. We found that this translation does not follow the Hungarian as precisely as the German text does. The hyphenation of the songs does not follow the original style, the distribution of the hapax legomena is different in this compared to the other two, and finally the texts are divided into chapters in a rather arbitrary way, definitely not following the original method. This last fact and the results gained again give a proof that there is no direct connection between an increase in the number of the newly introduced word-types and the physical appearance of a new chapter.

5. Summary

Based on a previously developed theoretical background, namely, that the vocabulary size and richness of literary works can be modelled using the randomness assumption, several models have been brought to life. The best results were obtained assuming that the selection of the words of the texts follows a hypergeometric distribution. Using this assumption we built a dynamic model which is able to imitate the text in progress, to give details about the appearance of the word-types from the beginning to the last words of the texts, unlike the methods mentioned earlier, which try to describe the overall vocabulary size and the vocabulary richness.

Using our model, based on the frequency and the relative frequency of the word-types, artificial texts can be created. To follow the narrative and to trace the behaviour of the appearance of words, the number of the newly introduced word-types were counted and plotted in both the original and the artificial texts. As it was shown these artificial texts were able to follow the general trends of the original texts but not the seasonalities which produced protuberances on the graphs of the newly introduced word-types. Analyzing the original texts, these protuberances were found to occur when the narrative is interrupted by a longish text-slice which is different in style from the main stream. In previously published but much less objective works one can find indications which are in accordance with our findings but merely subjective opinions also can be found which state that the number of the newly introduced word-types rises at the beginning of a new chapter, or in case of concatenated short stories at the beginning of a new story.

As we have seen by comparing the original and the corresponding artificial texts those opinions were confirmed which state that the authors can deliberately change the flow of the narrative and then switch back to the original stream.

To give further proof of our hypothesis we applied a new method, the comparison of the original works to their translations. This method was able to strengthen our results gained from the comparison of the original texts and their corresponding artificial texts that models assuming the independent usage of words in a text differ from the original text not because of the syntactic and semantic constraints but due to changes that occur on discourse level.

References

Baayen, R. H. 1996a: „The randomness assumption in word frequency statistics", in Perissinotto, G. (ed); *Research in Humanities Computing 5*, Oxford University Press, 17-31.

Baayen R. H. 1996b: „The effect of lexical specialization on the growth curve of the vocabulary", *Computational Linguistics* 22: 455-480.

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

Balázs, J. 1985: *A szöveg,* Gondolat, Budapest

Carroll, J. B. 1967: „On sampling from a lognormal model of word frequency distribution", in Kucera, H. – Francis, W N. (eds): *Computational Analysis of Present-Day American English,* Providence: University Press of New England

Csernoch, M. 2004: „Another method to analyze the introduction of word-types in literary works and textbooks", *The 16th Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities,* Göteborg University

Csernoch, M. 2006: Frequency-based dynamic models for the analysis of English and Hungarian literary works and coursebooks for English as a second language; Teaching Mathematics and Computer Science, Debrecen

Genette, G. 1980: *Narrative Discourse. An Essay in Method*, Cornell University Press, Ithaca/NY

Holmes, D. I. 1994: „Authorship attribution", *Computers and the Humanities* 28: 87-106.

Hoover D. L. 2003: „Another perspective on vocabulary richness", *Computers and the Humanities* 37: 151-178.

Mandelbrot, B. 1962: „On the theory of word frequencies and on related Markovian models of discourse", in Jakobson, R. (ed): *Structure of Language and its Mathematical Aspects,* Providence: University Press of New England

Muller C. 1964 « Calcul des probabilités et Calcul d'un vocabulaire », in *Travaux de linguistique et de littérature*, 235-244.

Scarborough, J. B. 1966: *Numerical Mathematical Analysis*, Johns Hopkins Press, Baltimore

Sichel, H. S. 1986: „Word frequency distributions and type-token characteristics", *Mathematical Scientist* 11: 45-72.

***

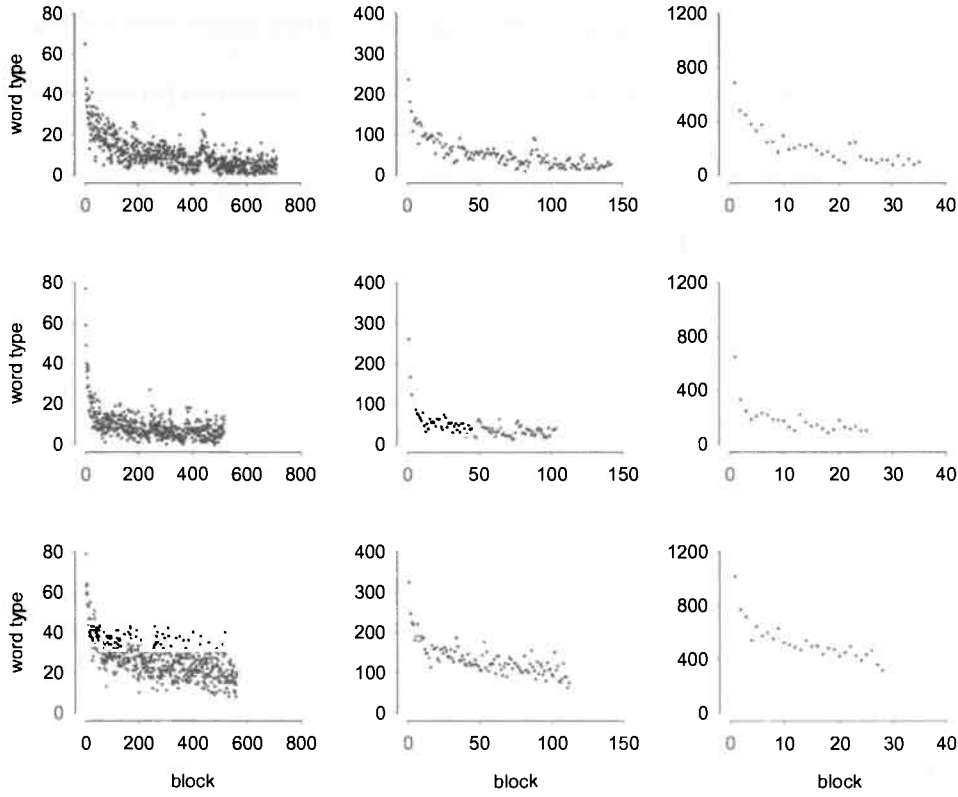

Figure 1. The effect of increasing the size of the block on the graph of newly introduced word-types. The calculation was done using one hundred- (left), five hundred- (middle), and two thousand- (right column) token long blocks on *Tom Sawyer* (upper), *The Jungle Book* (middle), and *Sorstalanság* (lower row). Note that together with decreasing the noise the use of larger blocks tended to eliminate secondary, small humps.

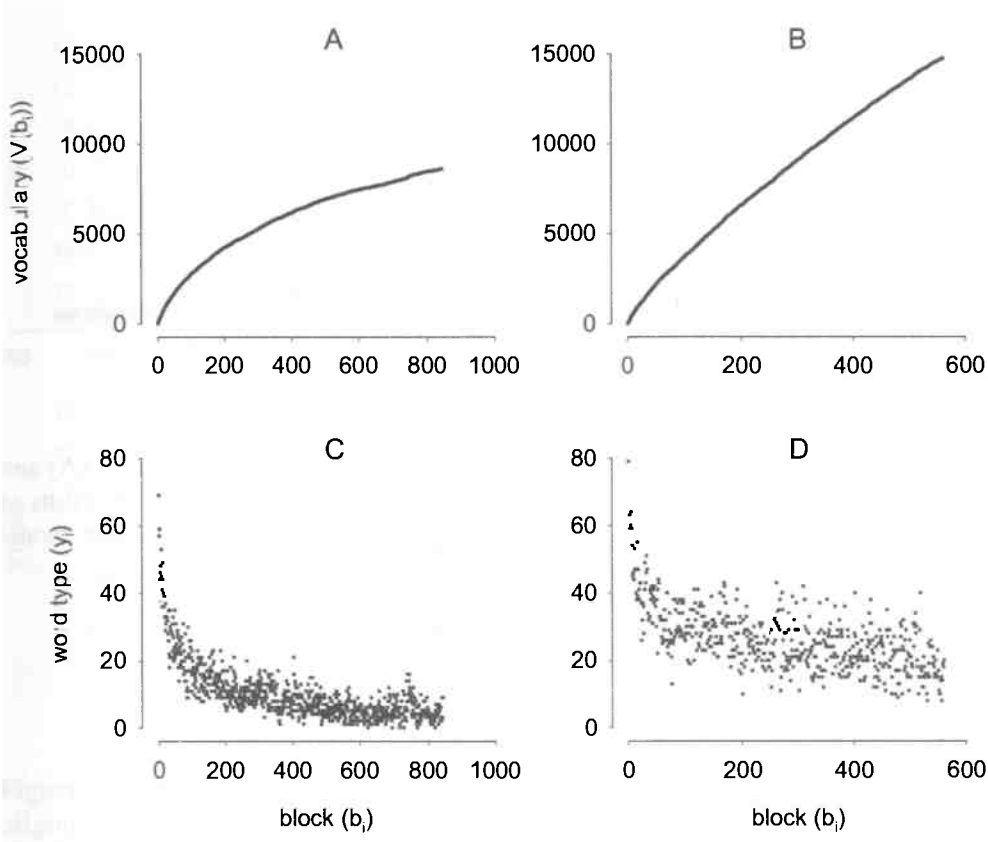

Figure 2. The cumulative vocabulary size ($V_{(bi)}$) (panel A and B) and the number of the newly introduced word-types (y_i) (panel C and D) in an English and in a Hungarian novel of similar lengths. The English novel is the *Scarlet Letter* (panel A and C), while the Hungarian is *Sorstalanság* (panel B and D). As it was expected, due to the characteristics of the Hungarian language, that it is morphologically rich (Table 2 and 3), the number of the newly introduced word-types in a block and consequently the vocabulary size is larger in a Hungarian text then in an English text of similar lengths (Table 4.).

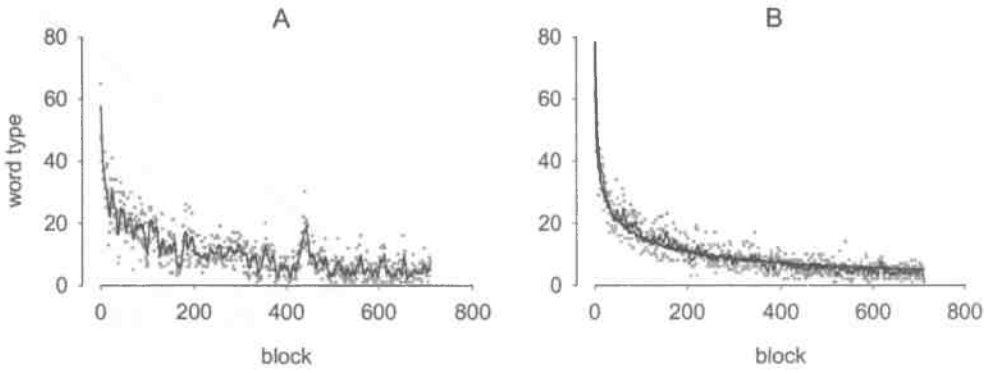

Figure 3. The number of newly introduced word-types (y_i) in *Tom Sawyer* (A) and its model prediction (B). Superimposed on the original data (dots) are the results of a 7-point smoothing (continuous lines in A). For comparison the average model (model ran one hundred times and averaged) is also shown in B.

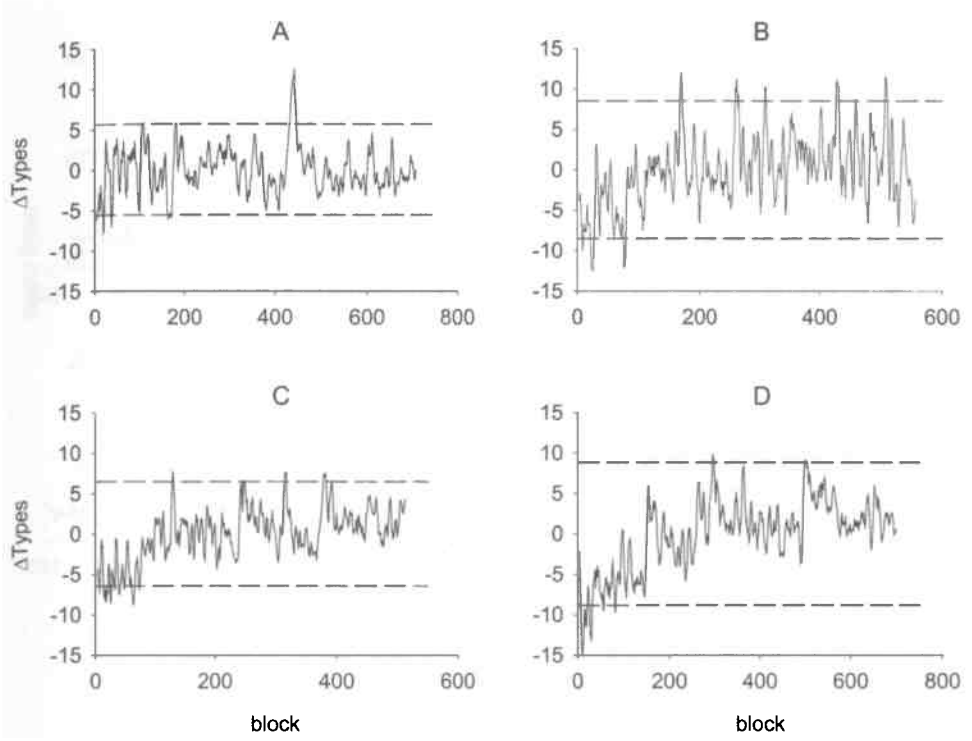

Figure 4. Differences in the number of newly introduced word-types between the original text and the model. The original text was filtered and the average model (model ran one hundred times and averaged) subtracted. Dashed lines mark mean ± 2σ calculated from the difference curves. Novels analyzed are *Tom Sawyer* (A), *Sorstalanság* (B), *The Jungle Book* (C) and *American Mystery Stories* (D).

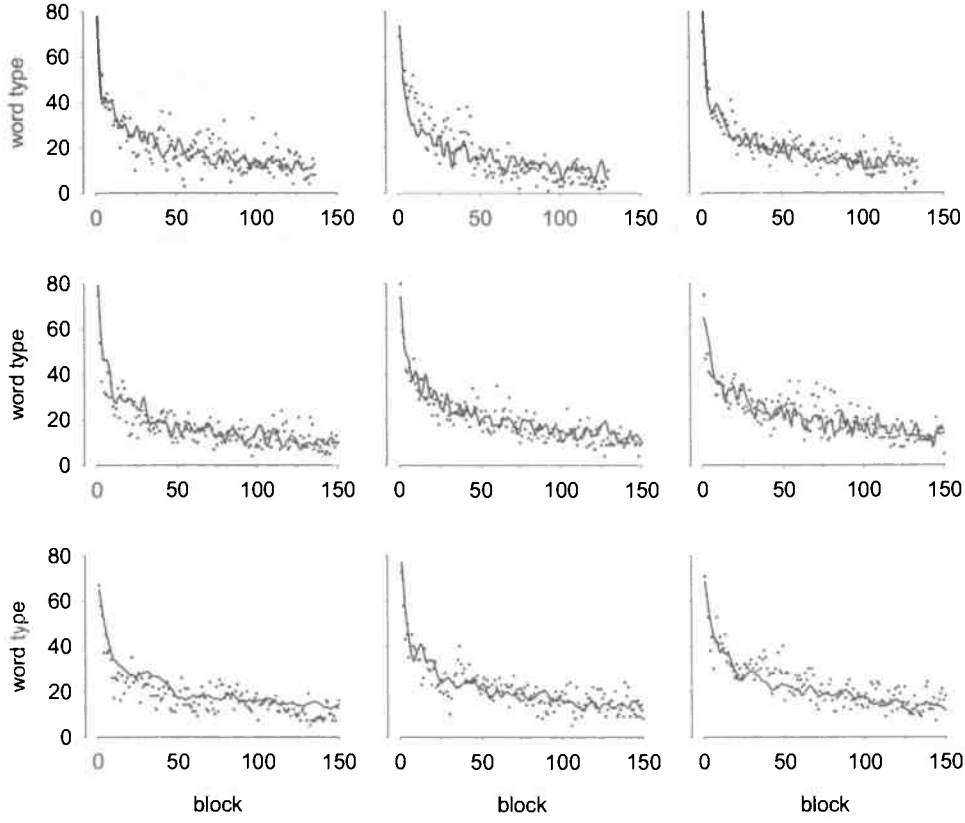

Figure 5. The comparison of the first one hundred and fifty blocks of English literary works of different lengths: short stories in the top row, middle-length-texts in the middle row, and long texts in the bottom row. The works in the middle column are from Jack London: *An Odyssey of the North, Call of the Wild, Seawolf.* In the right column they are from Joseph Conrad: *Youth, The End of the Tether*, and *Lord Jim*, while in the left column works are from different authors: The *Gold-Bug* from Edgar Allan Poe, *Harry Potter and the Philosopher's Stone* from J. K. Rowling, and finally *Great Expectations* from Charles Dickens. As it was earlier predicted and shown, neither the overall length of the text nor the author influence the vocabulary size significantly. This figure with nine panels shows that there are no distinguishing differences in the methods how word-types are introduced, also. The way the word-types are introduced is regardless of either the overall length or the author of the text.

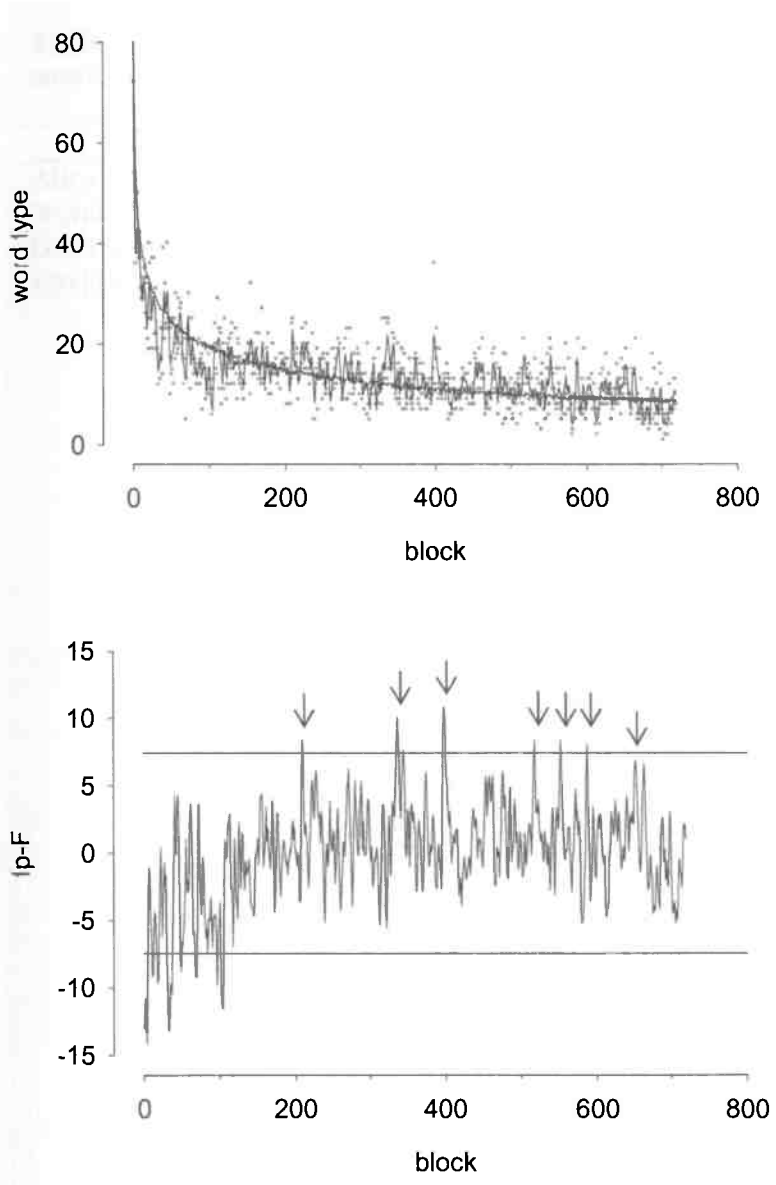

Figure 6. The German translation of „Sorstalanság": *Roman eines Schicksallosen.* Apart from the small differences the protuberances in the newly introduced word-types were found at the same locations in the Hungarian and German texts. Arrows mark the protuberances of the German text.

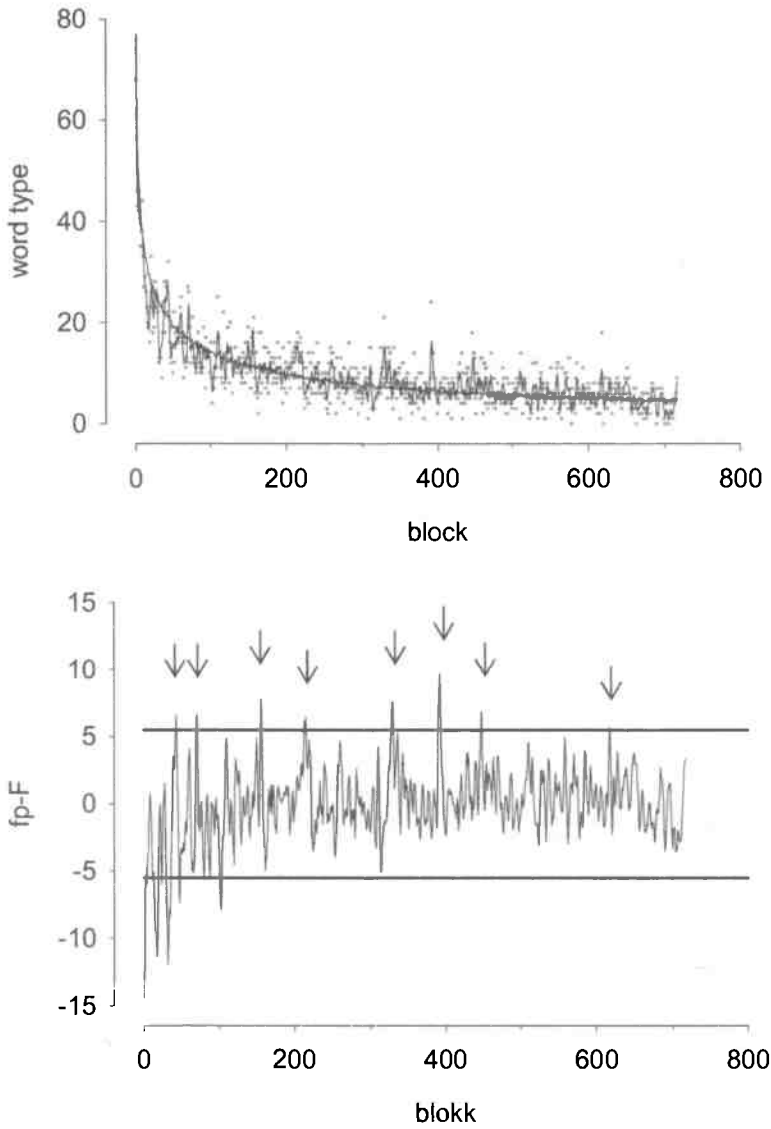

Figure 7. The protuberances on the graph of the number of the newly introduced word-types of *Fateless*, the English translation of „Sorstalanság".

***

Table 1. The number of tokens (N) in a text and in a block (h) using 20 equally spaced blocks.

	N	h
Alice's Adventures in Wonderland	26600	1330
Great Expectations	186500	9325
David Copperfield	358000	17900

Table 2. A comparison of how the affixation rules change the number of word-types in English and Hungarian texts. In the first and second columns the affixes are in bold to indicate which is added newly. In the third and fourth columns the cumulative vocabulary sizes are shown. In English the affixes previously used do not increase the number of the newly introduced word-types and so do not the cumulative vocabulary size.

Hungarian	**English**	**Hungarian**	**English**
ház	house	1	1
ház**ak**	house**s**	2	2
házak**ban**	**in** the houses	3	1+2
házak**ból**	**from** the houses	4	2+2
ház**aink**ból	from **our** houses	5	3+2
ablak	window	6	3+3
ablak**ok**	window**s**	7	3+4
ablakok**ban**	**in** the windows	8	3+4
ablakok**ból**	**from** the windows	9	3+4
ablak**aink**ból	from **our** windows	10	3+4

Articles were not counted.

Table 3. The vocabulary size ($V(N)$) and the number of hapax legomena ($V(1,N)$) in English and Hungarian texts of similar lengths, where n is the number of the blocks using hundred-token-long blocks. (H is for Hungarian and E is for English.)

		n	V(N)	V(1,N)
Ábel a rengetegben	H	517	11571	7856
The Jungle Book	E	516	4688	2064
Sorstalanság	H	561	14740	10253
Alice Adventures in Wonderland	E	562	3879	1515
Egri csillagok	H	1916	16237	8938
Harry Potter 4.	E	1920	10666	4298
A kőszívű ember fiai	H	2037	30677	21298
Great Expectations	E	1865	11022	4751

Table 4. Correct Hungarian sentences using the same four* words and their English translations. In the Hungarian sentences the same four words are used with different word order.

Hungarian	**English**
Tegnap mindenki eljött hozzánk.	Yesterday everyone came to our place.
Tegnap jött el hozzánk mindenki.	It was yesterday that everyone came to our place.
Tegnap hozzánk jött el mindenki.	It was our place that everyone came to yesterday.
Mindenki eljött hozzánk tegnap.	Everyone came to our place yesterday.
Hozzánk jött el mindenki tegnap.	It was our place that everyone came to yesterday.
Eljött hozzánk mindenki tegnap.	Everyone did come to our place yesterday.
Eljött hozzánk mindenki tegnap?	Did everyone come to our place yesterday?
etc.	

*In Hungarian the affixes to verbs can either serve as prefixes or suffixes with different syntax. If the prefix is in front of the verb the prefix and the verb make one word, if it is after the verb and the suffix are split up (e.g. *eljött* and *jött el*, where *el* is the affix to the verb *jött*).