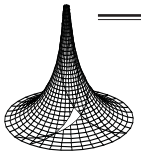

4/1 (2006), 53–70
tmcs@inf.unideb.hu
http://tmcs.math.klte.hu

Teaching
Mathematics and
Computer Science

Frequency-based dynamic models for the analysis of English and Hungarian literary works and coursebooks for English as a second language

Mária Csernoch

Abstract. We examined the characteristics of how word types are introduced in English and Hungarian literary works as well as in English coursebooks written for second language learners. By subdividing the texts into small segments of equal length, we were able to pinpoint subtle changes in the narrative. Based on the frequency of the word types in the original text a model was generated, and applying the model artificial texts were created. By comparing the original and the artificial texts, the places where these changes within the narrative occurred, could be located. Studying coursebooks we found that their vocabulary and how they introduce word types resembled those of randomly collected and concatenated short stories. According to our observations writers of the coursebooks forget that not only should the number of word types be carefully planned, but their repetition, in sufficient number, should also be cared for.

Key words and phrases: literary works, text analysis, text modeling, coursebooks, vocabulary analysis.

ZDM Subject Classification: K40, M80.

Introduction

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 [6]). The best results were obtained

with models that assume that word types[1] 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.

If ω_i $(i = 1, \ldots, V(N))$ mark the frequency of $f(i, N)$, the ith word type in the frequency order of an N token long text, then the appearances of the word types can be modelled with the multinomial distribution in the following way.

Let $A_1, \ldots, A_{V(N)}$ be a random vector, a set of random variables, with $p_i = P(A_i) > 0$, $i = 1, \ldots, V(N)$. If we assume that we have N independent trials $(\sum_{i=1}^{V(N)} p_i = 1)$, and ω_i marks the number of the outcomes of the A_i event, then the $(\omega_1, \ldots, \omega_{V(N)})$ joint distribution is an N and $(p_1, \ldots, p_{V(N)})$ parametric multinomial distribution:

$$\omega_1 = k_1,\, \omega_2 = k_2,\, \ldots,\, \omega_{V(N)} = k_{V(N)}, \quad k_1 + k_2 + \cdots + k_{V(N)} = N, \tag{1}$$

$$P\{\omega_1 = k_1,\, \omega_2 = k_2,\, \ldots,\, \omega_{V(N)-1} = k_{V(N)-1},\, \omega_{V(N)} = k_{N-(k_1+\ldots+k_{V(N)-1})}\} = \\ = \frac{N!}{k_1! \cdots k_{V(N)-1}!\,(N - k_{V(N)})!}\, p_1^{k_1} \cdots p_{V(N)-1}^{k_{V(N)-1}}\, p_{V(N)}^{N-(k_1+\ldots+k_{V(N)-1})}, \tag{2}$$

$$\sum \frac{N!}{k_1! \cdots k_{V(N)-1}!\,(N - k_{V(N)})!}\, p_1^{k_1} \cdots p_{V(N)-1}^{k_{V(N)-1}}\, p_{V(N)}^{N-(k_1+\ldots+k_{V(N)-1})} = 1. \tag{3}$$

In our case the trial is the selection of a word type from the text. If a word type is selected and marked as different from the others the multinomial distribution is reduced to the binomial distribution. Each of the k components separately has

[1]Tokens or running words are the strings of a text between two separator characters defined by the character set. Each unique instance of the tokens is a word type of the text.
Using these definitions of words *run*, *ran*, *runs*, *running* are counted as four word types of the same lemma (*run*). The hindrance of working with lemmas is that words which have the same orthographic form but different part-of-speech (POS) categories (the noun *run* and the verb *run*) are counted only once. To get more reliable information about the usage of words within a text the POS tagging has to be carried out without a simple lemmatization.
Considering all these it is obvious that both methods (counting the word types or counting the lemmas) in analyzing the words within a text have their pros and contras. One of the advantages of working with the forms of the words occurring in the texts (word types) is that they carry information about time, style, mode, relations, etc. in their inflections and thus provide clues for changes in the flow of text. On the other hand, in morphologically rich languages this method might cause extremely high number of word types based on the same lemma.

a binomial distribution with parameters N and p_i, for the appropriate value of the subscript i:

$$P\{\omega_{i_1} = k_{i_1},\, \omega_{i_2} + \ldots + \omega_{i_{V(N)-1}} = k_{N-(k_{i_2}+k_{i_3}+\ldots+k_{i_{V(N)-1}})}\} = \\ = \binom{N}{k_1} p_{i_1}^{i_1} (1-p_{i_1})^{N-(k_{i_2}+k_{i_3}+\ldots+k_{i_{V(N)-1}})} \tag{4}$$

To further simplify the presentation we can use the urn model to explain the problem and draw a comparison between classical statistics and our special linguistic problem.

It is assumed that an urn contains S different word types ω_i, $i = 1, 2, 3, \ldots, S$. With each word, ω_i, a population probability is associated, $\Pr(\omega_i) = p_i$, p_i, $i = 1, 2, 3, \ldots, S$. These probabilities can be thought of as the number of marbles with color i in an urn, divided by the total number of marbles in the urn. Sampling a word consists of randomly selecting – i.e., the selections are independent and have the same probability distribution $\Pr(\omega_i) = p_i$ – a word token from the urn, inspecting its color, and returning it to the urn. Because they are sampled with replacement, the probabilities of the words do not change over time. If $f(i, N)$ is the frequency of ω_i in a sample of N tokens, the probability ($Pr(f(i, N) = m)$ that ω_i appears exactly m times in a sample of N tokens was counted in the following way. We can consider the sample of N tokens as a sequence of N trials with m success (ω_i was drawn) and $N - m$ failures (ω_i was not drawn). The probability – assuming the independency – of a particular sequence of m successes and $N - m$ failures equals $p_i^m (1 - p_i)^{N-m}$. How many such sequences are there? This question can be rephrased as: In how many ways can we select m trials from N trials to be labeled as a success? The number of objects from N objects is the number of combinations of N objects taken m at a time: $\binom{N}{m}$. Hence, $\Pr(f(i, N) = m)$ equals

$$\Pr(f(i, N) = m) = \binom{N}{m} p_i^m (1 - p_i)^{N-m}, \quad \sum_{m=0}^{N} \binom{N}{m} p_i^m (1 - p_i)^{N-m} = 1. \tag{5}$$

This probability distribution is, again, called binomial.

Given the urn model, the frequency of a word ω_i with the probability p_i in a sample of N tokens is binomially (N, p_i) distributed. The expected frequency of ω_i (mean value) in the sample:

$$E[f(i, N)] = N \cdot p_i. \tag{6}$$

If $V(m,N) = \sum_{i=1}^{V(N)} I_{[f(i,N)=m]}$: the number of types with frequency m in a sample of N tokens, the expected number of word types with frequency m in a sample of N:

$$E[V(m,N)] = \sum_{i=1}^{S} \binom{N}{m} p_i^m (1-p_i)^{N-m}. \tag{7}$$

The expected number of different word types in the sample is

$$E[V(N)] = E\left[\sum_{m=1}^{N} V(m,N)\right] = \\ = \sum_{m=1}^{N}\sum_{i=1}^{S} \binom{N}{m} p_i^m (1-p_i)^{N-m} = S - \sum_{i=1}^{S} (1-p_i)^N. \tag{8}$$

The ultimate aim of the present work is to build a dynamic model in order to understand the underlying reasons for the introduction of new word types in texts or, in other words, to examine the circumstances which lead authors to introduce a new word type. To carry out the experiments, a unique model was built for each text, written in any natural language. Based on this model, artificial texts were created with the constraint that the original and artificial texts must have the same frequency for each word type. By comparing the original and the artificial texts quantitative data could be deduced and used to find reasons for the authors' strategies in using different word types in a text.

Methods

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

$$frel(i,N) = \frac{f(i,N)}{N} \in \left]0;1\right[, \tag{9}$$

thus, the probability of occurrences (p_i). While previous works focused on the overall vocabulary size ($V(N)$) and vocabulary richness the given formulae were able to produce reliable pieces of information. However, our aim was not the determination of the vocabulary size but 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 its progress. Given these constraints, new methods with a new theoretical background had to be found.

The essence of our method is to create artificial texts using the frequencies and relative frequencies of the word types of the original text.

To start creating an artificial text, first the size of its vocabulary had to be determined. This is not surprising since writers do the same. They have an actual vocabulary, whose size is continuously changing [16]. Only for some writers is their total work digitalized and analyzed thoroughly, but even knowing the number of words used in their works does not guarantee that we have any idea about the size of the writers concerned. Especially, if we consider, on the one hand, that one's own vocabulary is an ever growing set of words [15, 16] and, on the other hand that it is difficult to tell the receptive and the productive (passive/active) vocabularies apart. Therefore, the words are picked from this huge but indeterminable vocabulary in a hitherto unfamiliar way. The present approach might help to trace any sign of the writers' strategy.

Based on the relative frequencies of the word types, a distribution function (*Femp*) is generated for each original text, where each word type (ω_i) is represented with its own relative frequency ($frel(i, N)$).

$$Femp(j) = \sum_{i=1}^{j} frel(i, N), \quad j = 1, \ldots, S. \tag{10}$$

Randomly selecting numbers from the $]0; 1[$ interval and mapping them to the word types through the distribution function allows the generation of randomly selected words which have the same probability of occurrence as in the original text. This random selection is repeated until the number of words in the model text reaches that of the original. With this simple method model texts can be generated in which the probability of a given word type equals that of the original text.

There is, however, a slight problem with the above algorithm. Since the word types are selected randomly, that is only their frequency is set, there is no guarantee that each and every word type will actually appear in the generated text. Indeed, running the program repeatedly gave, as expected, consistently smaller numbers of word types in the generated than in the original text (Figure 1). The discrepancy was the largest for words that appear only once (hapax legomena) in the original text. In order to correct this slight difference between the original and the generated text the algorithm was modified by artificially increasing the number of word types from which the random selection was carried out. In order not to change the frequency of all word types the following strategy was implemented. The number of hapax legomena was increased so that the relative

frequency and the probability of each hapax legomenon were decreased. This was carried out with the constraint that the overall relative frequency of hapax legomena should not be changed.

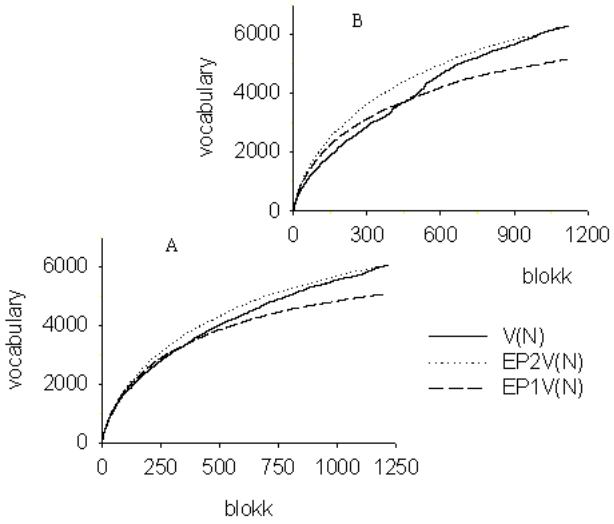

Figure 1. The graphs show the vocabulary size of two middle size corpora (A: Daniel Defoe: The Adventures Robinson Crusoe, B: Mark Twain: The Adventures of Huckleberry Finn). The continuous lines are for the word types of the original text ($V(N)$), the dashed lines show the generated word types with the original algorithm ($EP1V(N)$), while the dotted lines show the result obtained with the modified algorithm ($EP2(N)$).

The relative frequency of hapax legomena in an N token long text is

$$rel(V(1,N)) = \frac{V(1,N)}{N}. \tag{11}$$

If we use this equation, the relative frequency of a new word type becomes

$$x = \frac{V(1,N)}{N(V(1,N)+V2)}, \tag{12}$$

where $V2$ is the number of the newly added word types.

In the static model of Baayen [2, 3, 4, 5, 6, 13] the texts are divided into 20 or 40 segments of equal length, therefore, the calculations and the graphical presentation can be carried out in 20 or 40 different points only. Since their

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

aims were to examine vocabulary size and vocabulary richness, which values are independent of the length of the segments, a constant for the number of the segments, in spite of the fact that it provides text slices of different lengths for texts of different lengths, was a reasonable choice.

In contrast, 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 segments in question, segments of different lengths could not be used. Considering all these, our model differs from those presented earlier in that that the texts are not divided into an equal number of segments independent of the length of the given text. Instead, we kept the lengths of the blocks constant (h). To carry out this new method a suitable constant for the length of the segments had to be chosen.

Usually blocks containing one hundred tokens ($h = 100$) are chosen. Therefore, the number of blocks varies from text to text. 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 slices from different texts can be readily compared. A shorter and a longer text divided into 20 or 40 equally spaced segments are not comparable considering either the number of tokens or the word types.

As Table 1 shows, using the previously published method for slicing up the texts provided in David Copperfield one gets almost as long blocks as the whole length of Alice's Adventures in Wonderland, which result indicates that the length of the blocks, for our purposes, had to be chosen by using a different method.

The second advantage of using constant-length-long blocks comes from the relatively short length of the blocks. Using these short blocks subtle changes, couple of hundred-token-long text slices, in the narrative can also be traced (Figure 5).

To show the effect of longer than one hundred-token-long segments we picked texts of different lengths divided into 40 equally spaced blocks. Figure 2/A and C show that there are details of the texts marked by a relatively high number of newly introduced word types which are overlooked by the choice of longer blocks (Figure 2/B and D).

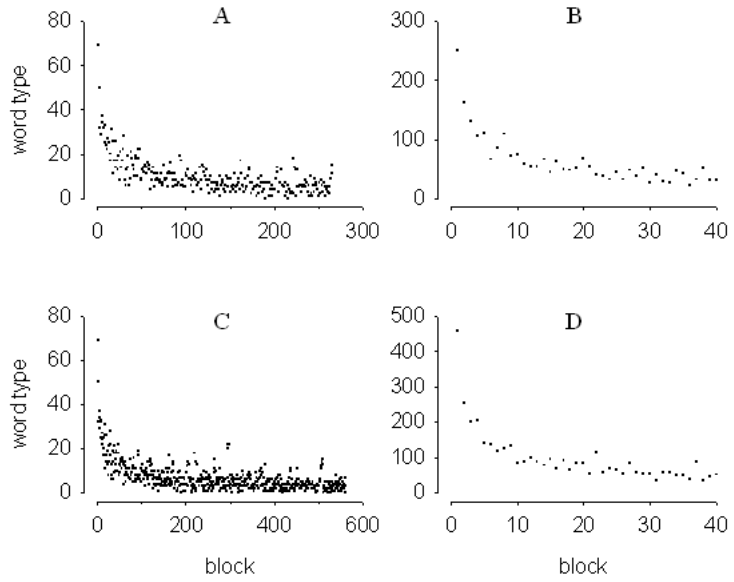

Figure 2. Lewis Caroll: Alice's Adventures in Wonderland and Through the Looking Glass. A and B graph the number of the newly introduced word types in the sequential blocks in Alice's Adventures in Wonderland while C and D show the concatenated texts of two Alice stories. In A and C a hundred-token-long blocks are used, while in B and D the texts are divided into 40 equally spaced blocks. In the longer text many of the peaks of the shorter text are diminished.

Materials

Although the "word" is clearly central to understanding and analysing a language, one would look in vain for a simple definition of the concept "word". On the one hand, words can be thought of in terms of types and tokens; *Tomorrow,*

and tomorrow, and tomorrow will be thought of as containing five words (*tomorrow, and, tomorrow, and, tomorrow*) or two words (*tomorrow, and*) depending on whether one is viewing words as tokens (actual occurrences of any item) or types (items with different identities). Likewise, the phrase *Going, going, gone* will be considered to comprise three words (*going, going, gone*) on a count of tokens but only two words (*going, gone*) on a count of types [14], [16]. According to this usage, word tokens and types in texts have been counted and examined in our experiments. Still, to find any of these in a text first a character set had to be set up.

To carry out the experiments a piece of software, DyMoCASAT (Dynamic Models for Computer Aided Statistical Analysis of Texts) was developed. DyMoCASAT carries out the data retrieval from the original text and the building of the model, and – based on the model – the generation of the 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.) The English set contains both the twenty-six lower- and uppercase letters of the English alphabet and the apostrophe, while the Hungarian uses both the thirty-five single-character lower- and uppercase letters. According to this method of determining the size of the character set, the cardinality of the English character set is $c = 2 \cdot 26 + 1$, while the Hungarian is $c = 2 \cdot 35$. Any other character which is not in the defined character set, such as space, comma, sentence-closing marks, numbers etc. and the apostrophe in Hungarian texts serves as a separator.

For the analyses we needed the electronic version of the original, printed texts. The main source for these electronic versions was the Internet. The texts that were not available for 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.

Before the final processing, the texts from different sources had to be standardized and formatted. To get results comparable with previous ones neither standardizing nor formatting meant preprocessing of the texts. Instead, a filtering had been carried out where we had to correct the typing and scanning mistakes, deal with and somehow unify the different typographic conventions, deleting those paragraphs of the e-texts which were added to the original works, concatenate the units, chapters saved separately, and finally convert them into text format and save them, since the program opens and works with unformatted text files.

Results

Data retrieval from literary works

DyMoCASAT, as the first step of the analysing process, saves all the possible information about the location of each word of the selected text in text (`.txt`) files (by default in the `Windows\Temp` folder).

Figure 3. The word types of a text are stored in text files, named after the ASCII codes of their initial characters. From the size of the file we can conclude the number of words starting with that particular character and further on the frequency of the appearance of that particular character as the first letter of words in a language.

- The number of the files equals the number of the initial characters of the word types of the text (Figure 3). This number is usually equal to the cardinality of the character set, however there are cases, especially in short stories, when the number of the files does not reach it. In such cases text files which belong to rare characters are not created.
- Each file contains all the words starting with that particular letter (character). The words are kept in separate paragraphs, so each word has its own paragraph (Figure 4).
- Each paragraph, opening with the word, carries at most as many characters as the number of blocks into which the text is divided ($n = \left\lceil \frac{N}{h} \right\rceil$; Figure 4).

The content of these paragraphs are ASCII-characters. The character at the ith position indicates the number of occurrences (c_i) of the corresponding word in the ith block, using the notation that c_i equals the ASCII-code of the character minus 64 (e.g. the @ character denotes that no corresponding word was present in that block).

```
the
FFDCCFECBBFGFDDACDFAEGCFCFJFDG@BFEEADDDDBDFEDIIGGFHEFFFDGDDDFGDDFDFBACGECBE
EDDFEFDHEDGHCBGCEHHECGFEABFEFCCEADEMDFHGJDFFGEKEDDACCDDDGEJFEDDHIJGJCEIGHFD
FELGEEGLCCFIEDHKGHFHCJCDCIGHHHGDNBBDIHFBFGKHIHEDEFGHDHHHDJLHGIFKDFEFGHEIJJD
KIFIGEGNJMFDNNJHKQHKKGGICGGLHFDEEHLFAEMPO
their
@@@@@@A@@@@@@@@A@@@@@@@@@@@@@@BA@@@@@@@@@@@A@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@AAB@@@@@@@@A@@@@@@@@@@@@@@@A@@@@@@@@@@@@A@
@@@@@@@@@@AA@AB@AB@@A@@@@@@@A@A@@@@@A@@@@@A@@@@@@@@@@@B@@@@A@@@DB@@A@@@@@@@
@A@@AC@@@A@@@@@@@@@@@AAAA@@@@@@A@@A@@@A
theirs
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@A
```

Figure 4. The appearance of *the*, *their*, *theirs* from Lewis Caroll: ALICE'S ADVENTURES IN WONDERLAND (a fragment from file `116.txt`). Since *the* is the most frequent ($f(1, N) = 1639$, $N = 26600$) word, it is not surprising that it occurs in each block.

Based on the stored data several further values were calculated and saved by DyMoCASAT in an N-token-long text:

- the number of word types ($V(N)$),
- the frequency of each type ($f(i, N)$, $i = 1, \ldots, n$), and
- their relative frequency ($frel(i, N)$).

As mentioned earlier, DyMoCASAT analyses both English and Hungarian texts. The graphs of Figure 5 clearly demonstrate that the number of the newly introduced word types in the blocks is clearly higher in Hungarian than in English texts and also is their cumulative value, the vocabulary size ($V(N)$). This, however, does not necessarily reflect a larger vocabulary, rather it is due to the agglutinating characteristic of the Hungarian language. To enable the comparison of the vocabulary of the two languages, as well as the vocabulary of a text and its translation, morphological analyser(s) and part-of-speech tagger(s) would be needed [7] and [8]. Since our aim is not the analysis of either of these problems, the raw texts, without any preprocessing, are analyzed. Aware of previously published [11] but rather subjective intuitions that inflections also carry reliable pieces of information, we kept the inflections at their original places.

Trends in the appearance of 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 ($f(b_i) = y_i$, $i = 1, \ldots, n$; e.g., Figure 5). The number of newly introduced word types, in general, follows

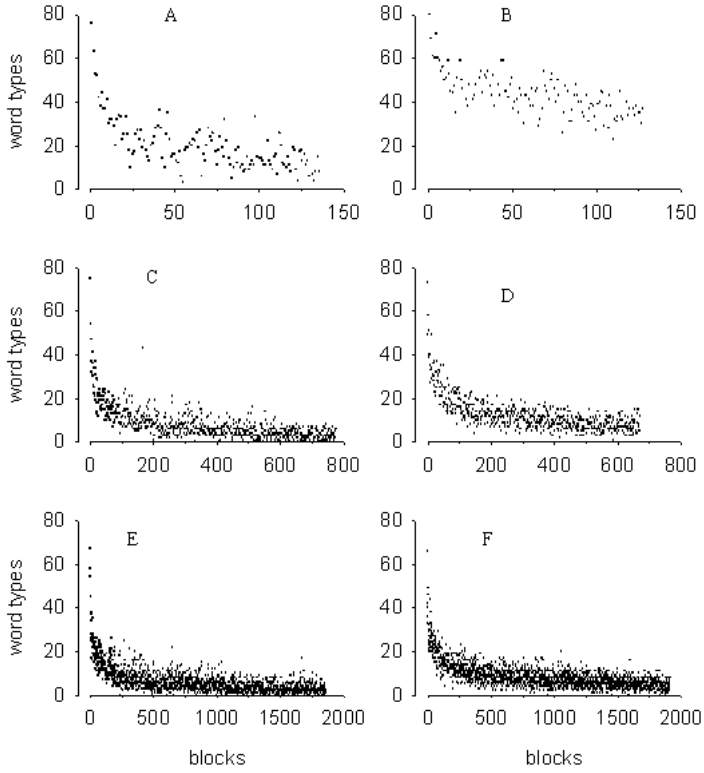

Figure 5. The introduction of word types in English (left) and Hungarian (right) texts of different lengths. The texts are divided into a hundred-word (token) long blocks. The graphs show the number of newly introduced types in each block in texts with different lengths. At the top results from "short", approximately 15,000-token-long-texts, in the middle from "middle size", approximately 80,000-token-long-texts, and at the bottom from "long", approximately 150,000-token-long texts are shown. (A: Edgar Allan Poe: The Gold-Bug, B: Rejtő Jenő: Vissza a pokolba!, C: J. K. Rowling: Harry Potter and the Sorcerer's Stone, D: Zsoldos Péter: A feladat, E: Charles Dickens: Great Expectations, F: Gárdonyi Géza: Egri csillagok.)

a decaying tendency. There are, however, parts of the texts where their number is greater than what is expected from this general trend. A point or a group of points that fall significantly outside the general trend form a local maximum in the neighbouring blocks, which is 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.

DyMoCASAT enabled us to locate those places of the original texts, where these protuberances occurred. Examining the narrative at these locations we found that the protuberances on the graphs appeared when a new character or place was introduced, a long description of an event interrupted the flow of the story, the author made such a character speak whose style and vocabulary are significantly different from those who have spoken before, and, finally, when foreign words, expressions or sentences were mixed into an otherwise monolingual text.

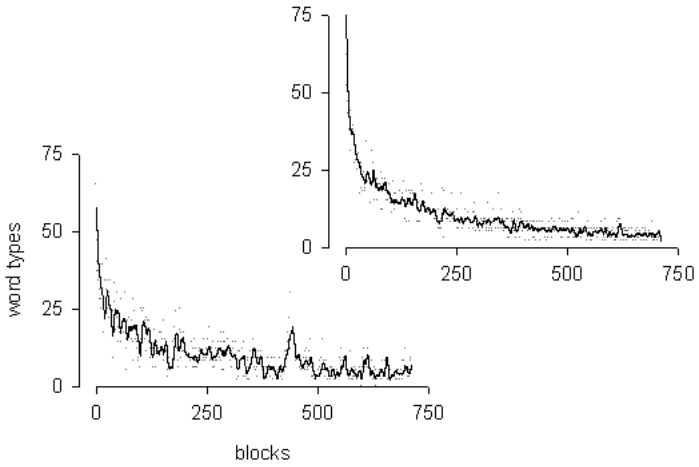

Figure 6. Mark Twain: The Adventures of Tom Sawyer. On the left the dots show the word types of the original text in each block ($h = 100$) while the line is a seven-point smoothing. On the right the word types of the artificial text and the smoothed graph based on these points are plotted. Comparing the two smoothed graphs the parts where the author has inserted a longish, relatively new topic into the flow of the text, marked by a protuberance, are recognizable.

The model generated by DyMoCASAT was capable of reproducing both the general, declining trend and the small fluctuations, appearing as noise or as a renewal process on the graphs of the newly introduced word types (Figure 6). On the other hand, events that did not fit naturally into the narrative or were not part of the main stream of events and appeared as secondary rising phases in the graphs, the protuberances, were not reproduced by the model (Figure 6).

Coursebooks for Second Language Learners

Not only literary works, but monolingual language coursebooks were also analysed. How these coursebooks should be composed has a vast literature in the methodology of language teaching [1, 9, 10, 12, 14, 15, etc.]. However, among the many aspects and criteria listed in these works the carefully planned selection of the vocabulary is one of the last. It is established that teachers should teach as many words as possible, or at least 1000 words per course (120–150 hours). The second criterion is that the most important and useful words should be taught. To fulfil this second criterion is not an easy task, since there are no unambiguous definitions for these expressions. Our aim was to use the above detailed method and DyMoCASAT to see what the differences were between literary works and coursebooks in the way they introduced word types.

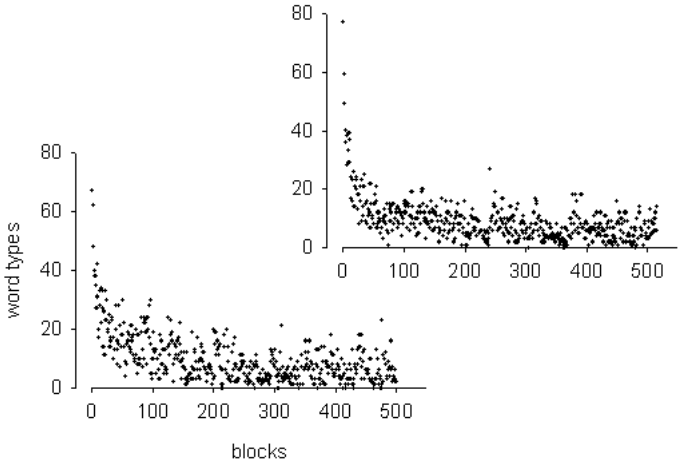

Figure 7. The introduction of word types in Kipling: The Jungle Books (left) and in Soars: New Headway Intermediate (right). We found that the introduction of word types in the English coursebook resembles that of concatenated short stories and of novels with similar lengths. Note that the number of word types is kept as high as in short stories.

The analysis of the selected series of coursebooks has shown that the appearance of word types in a language coursebook is not significantly different from that in concatenated short stories and novels with similar length (Figure 7, Table 2). Furthermore, it should be pointed out that the number of hapax legomena

is surprisingly high in these coursebooks (Table 3 and 4) and those teachers who use this book should carefully plan their work to be effective.

Table 2. The number of tokens (N), word types ($V(N)$), the average relative frequency ($V(N)/N$), and hapax legomena ($V(1,N)$) in a one-volume concatenated short stories and in the NEW HEADWAY INTERMEDIATE. The numbers show clearly that there is hardly any difference between the two books.

Author, title	block $n = \left\lceil \frac{N}{h} \right\rceil$	vocabulary $V(N)$	$\frac{V(N)}{N} \cdot 100$	hapax legomena $V(1,N)$
Kipling: The Jungle Book	516	4688	9.085	2067
Soars: New Headway Intermediate	500	4803	9.606	2072

Table 3. The number of blocks (n), types ($V(N)$) and their average relative frequency, and the number of hapax legomena ($V(1,N)$) in the NEW HEADWAY coursebooks. Up to the INTERMEDIATE book the number of tokens increases continuously but surprisingly after that there is not much difference in the number of tokens in the three most advanced books of the series. The other notable result is the high number of hapax legomena.

New Headway	n	$V(N)$	$\frac{V(N)}{N} \cdot 100$	$V(1,N)$
Beginner	163	1539	9.442	501
Elementary	239	2452	10.259	864
Pre-Intermediate	317	3309	10.438	1373
Intermediate	500	4803	9.606	2072
Upper-Intermediate	511	5646	11.049	2430
Advanced	513	6724	13.107	3274

The number of word types in the coursebooks is much higher than it was predicted [9, 10, 14] and, as Table 2 shows, not less than in a much less planned literary work. Unfortunately, all this is in accordance with previous suggestions that the approach taken to the vocabulary has not been systematic and that there has been little coordination in establishing targets [9, 14]. If we consider that a coursebook is suitable for an approximately 120 hours of study and an average of

eight to twelve productive items as representing a reasonable input [10] we can see that the numbers of the word types of the selected series are much higher. These high numbers make us think further if we consider that students seem unable to master this number of words in coursebook, even after the teacher's explanations and drills [1].

Table 4. The number of tokens, types, and hapax legomena in the concatenated New Headway coursebooks. The concatenated books are able to show how many word types the students have come across and are supposed to be familiar with using this series. The table shows that especially the number of hapax legomena is extremely high.

New Headway	n	$V(N)$	$\frac{V(N)}{N} \cdot 100$	$V(1,N)$
Beginner	163	1539	9.442	501
Beginner → Elementary	402	2943	7.321	962
Beginner → Pre-Intermediate	719	4550	6.32	1628
Beginner → Intermediate	1220	6760	5.54	2607
Beginner → Upper-Intermediate	1731	8989	5.193	3458
Beginner → Advanced	2245	11648	5.188	4636

Summary

Against 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 follow the multinomial distribution or its reduction to the binomial distribution.

Applying this method 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 overall vocabulary size and vocabulary richness.

With the help of 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 artificial

texts. As 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 was interrupted by a longish text slice which was 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 can also find merely subjective opinions which state that the number of the newly introduced word types rises at the beginning of a new chapter or in the 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 have been confirmed which state that the authors can deliberately change the flow of the narrative and then switch back to the original stream.

To see how a much more carefully planned text behaves we chose to analyze monolingual English coursebooks designed for second language learners. Before building the model of these books they were compared with literary works. Contrary to our expectations, we found that there was hardly any difference between the appearance of the word types in these coursebooks and randomly selected literary works. The resemblance was the greatest to concatenated short stories. It was not only the rhythm of the introduction of word types but also the number of word types and the number of the hapax legomena that showed great similarities.

References

[1] V. F. Allen, *Techniques in Teaching Vocabulary*, Oxford University Press, Oxford, UK, 1983.

[2] R. H. Baayen, H. Halteren and F. Tweedie, Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution, *Literary and Linguistic Computing* **11**, no. 3 (1996), 121–131.

[3] R. H. Baayen, Statistical Models for Word Frequency Distributions: A Linguistic Evaluation, *Computers and the Humanities* **26** (1993), 347–363.

[4] R. H. Baayen, The Effect of Lexical Specialization on the Growth Curve of the Vocabulary, *Computational Linguistics* **22** (1996), 455–480.

[5] R. H. Baayen, The Randomness Assumption in Word Frequency Statistics, *Research in Humanities Computing 5*, Selected Papers from the ACH/ALLC Conference, University of California, Santa Barbara, August 1995 (1996), 17–31.

[6] R. H. Baayen, *Word Frequency Distributions*, Kluwer Academic Publishers, Dordrecht, Netherlands, 2001.

[7] D. Biber, S. Conrad and R. Reppen, *Corpus Linguistics. Investigating language structure and use*, Cambridge University Press, Cambridge, 1998.

[8] T. Brants, Internal and external tagsets in part-of-speech tagging, *Proceedings of Eurospeech*, Rhodes, Greece (1997).

[9] R. Carter and M. McCarthy, *Vocabulary in Language Teaching*, Longman Group UK, London and New York, 1991.

[10] A. Cunningsworth, *Choosing your Coursebook*, Heinemann, 1995.

[11] G. Genette, *Narrative Discourse. An Essay in Method*, Cornell University Press, Ithaca, NY, 1995.

[12] R. Grains and S. Redman, *Working with Words. A guide to teaching and learning vocabulary*, Cambridge University Press, Cambridge, UK, 1992.

[13] D. L. Hoover, Another Perspective on Vocabulary Richness, *Computers and the Humanities* **37** (2003), 151–178.

[14] P. Nation and R. Waring, Vocabulary size, text coverage and word lists, in: *Vocabulary: Description, acquisition, and pedagogy*, (N. Schmitt and M. McCarthy, eds.), Cambridge University Press, Cambridge, UK, 1997.

[15] N. Schmitt, *Vocabulary in Language Teaching*, Cambridge University Press, Cambridge, UK, 2000.

[16] D. Singleton, *Exploring the Second Language Mental Lexicon*, Cambridge University Press, Cambridge, UK, 1999.

[17] G. U. Yule, *The Statistical Study of Literary Vocabulary*, Cambridge University Press, Cambridge, UK, 1944.

MÁRIA CSERNOCH
UNIVERSITY OF DEBRECEN
FACULTY OF INFORMATICS
H–4010 DEBRECEN
P. O. BOX 12
HUNGARY

E-mail: mariacsernoch@hotmail.com

(Received June, 2005)