Proceedings of the International Multiconference on
Computer Science and Information Technology pp. 389–396
ISBN 978-83-60810-27-9
ISSN 1896-7094

The Role of the Newly Introduced Word Types in the Translations of Novels

Maria Csernoch
University of Debrecen,
Egyetem tér 1. Debrecen, H-4029, Hungary
Email: csernoch.maria@inf.unideb.hu

***Abstract*—The project detailed in the article is able to find the vocabulary rich segments of novels in different languages. The method used here takes into account the frequency of the words of the text, and based on this information we are able to create artificial texts with the same parameters. Since the original and the artificial texts share parameters they are comparable and we can find those segments of the original text which are richer in vocabulary then it is expected as compared to a random selection of the words. The advantages of finding these vocabulary rich segments of the text, beyond that they give an insight of the development of the vocabulary of a novel, is that in any translation, adaptation process it is a great advantage being familiar with these sections of the text.**

I. Introduction

THE PRIMARY goal of this project was to find a dynamic first-order statistical model [5] with the help of which we are able to find the vocabulary rich segments of the texts written in natural languages. By vocabulary rich segments we understand those sections of the text where the number of the newly introduced words is higher than it is predicted by the model. These sections of the text make their stand out by being richer in vocabulary than the main body of the text.

After finding such a model, and – based on the model – the text slices corresponding to the vocabulary rich segment of the text the question arose, what could be the reasons for their presence. I found several previously published works which venture to give explanations for this phenomenon but these opinions are subjective, mainly based on the intuitions of the reader. I was interested in finding more objective explanations, which are based on quantitative data, for what makes the author of a text to increase the number of newly introduced words so intensively that the process makes qualitative changes in the text.

Finally, my aim was to explore how the proportion of newly introduced words varies in different adaptations of the same text. The term 'adaptation' is used here to include both intralingual and interlingual adaptations, whether involving reduction in text size or not. This means that beyond the theoretical significance that we were able to find a method with the help of which we can separate the vocabulary rich segments from the main body of the text, the practical usage of the method is also remarkable. One of the advantages of this statistical analysis is that it is invariant to the fact that translators are free to use words, expressions, structures, different kind of techniques well known in translation studies. The only expectation here is that in a faithful translation the text segments in the target language should be as rich in vocabulary as they are in the original text. Both translators and critics are offered the location and the intensity of the vocabulary rich segments of a text. On the one hand, prior to the translators' effective work becoming familiar with the vocabulary rich text segments of the original work might help the translator(s) to pay more attention to these unique slices of text. They are unique in the sense that the author of the original text used richer vocabulary to call the readers' attention to them. As a result, we would be able to reach a more faithful translation. On the other hand, being familiar with the vocabulary rich text segments of both the original and the translated texts means that a new set of parameters is added to the tools of translation criticism. Here we would be able to provide more objective analysis of the translation(s), we would be able to tell how close a translation is to the original text in vocabulary, or decide which translation is closer to the original text.

II. Methods

A. Newly introduced words (word types and lemmas)

The analyses of texts highly depend on the notion of newly introduced word types. First the definition of newly introduce word types should be given.

Newly introduced word types have meaning in a closed text, at a certain point of that particular text. When the first appearance of the word is detected at a certain point of the text, the word is considered as newly introduced in that text.

Beyond that newly introduced word types are understood within a closed environment, unlike other word concepts newly introduced word type is relative. The title being newly introduced is only temporary, and the word is in the range of this concept until its second appearance in the text. The definition of newly introduced words contain that hapax legomena of the text are those words which never loose this title.

B. Dynamic first-order statistical models

To be able to follow the flow of a text instead of the previously used static models a dynamic model should have been created. The advantage of the dynamic model to the static models is that it gives data not only in general but in

the comparison to the original text it is able to follow the changes of the text.

The model belongs to the category of first-order statistical models [5] because it takes into consideration only the frequency of word types of the original text. A first-order statistical model can be created with the urn model. The urn model for word frequency distributions compared the use of words to the sampling of marbles, balls from an urn [3]. The essence of urn model is the following. Consider an urn containing marbles of various colors. Each color corresponds with a marble type. A particular color may be extremely rare, or it may be represented by a great many individual marbles, the marble tokens. We randomly draw N marbles from the urn, assuming that the outcome of a given trial is completely independent from the outcome of any other trial [17], [19], [27], [1], [2], [3]. In Baayen's opinion [2], [3]the urn model is responsible for the overestimation bias, which means that using a model, where the polynomial distribution of words are assumed, produces a larger size of vocabulary, specially in the first half of the text than the observed vocabulary.

To decrease the difference between the observed and the predicted vocabulary instead of assuming the polynomial distribution of the words, the hypergeometrical distribution of them should be applied. The difference between the two models seems to be minor, but assuming the hypergeometrical distribution of the words the overestimation is usually not longer than a couple of thousand words at the very beginning of the texts and later on the observed vocabulary fluctuate around the expected vocabulary.

If we assume that the word types of the texts can be modeled by the balls with different colors of an urn, and from each color as many balls are stored as the frequency of the word type than for polynomial distributions the balls are randomly selected from the urn and are returned after taking notes of its color [3], [7], [8]. The selection of balls is continued until we reach the number of tokens of the original text.

With the assumption of the polynomial distribution of the words, however, there is no clear guarantee that when reaching the desired number of tokens the number of different word types of the observed text equals the number of the different word types of the model.

The hypergeometric distribution of the words, on the other hand, can be modeled by an urn where the balls are not returned after inspecting its color. With word types the algorithm is the following. The number and the frequencies of the different word types of the original text are counted and then all the found word types are stored in a one-dimensional array as many times as their frequencies indicate. The size of the array equals the size of the text that is the number of the tokens in it. The random selection of a word type beyond inspecting it means the erasing of it from the array. Using this method the algorithm might slow down towards the end of the process, because as we advance in selecting the words the place of the erased words are reselected more often than at the beginning. To speed up the algorithm after selecting and deleting the word from the array the cell was not left on its original location, rather, the array was compressed by moving the elements forward. In the next step a new word was selected from this, one cell shorter, array. As a result, even to the longest novels I ever met the selection was carried out within two minutes. The great advantage of the hypergeometric urn model to the polynomial urn model that the number and frequency of the word types in the model equals to that of the original text. All this was carried out with DyMoCASAT (Dynamic Model for Computer Aided Analyses of Texts), a program designed for these special analyses).

C. Texts and their different adaptations

In this project primarily English and Hungarian novels were compared to their different adaptations. The selected adaptations were the human foreign language translations to English, French, German, and Hungarian, and the machine translations to English and Hungarian languages. Here the only publicly available English−Hungarian translator program [20] was used to create the translations and then the same method with DyMoCASAT was applied to these texts. Beyond these inter-language adaptations the lemmatized and non-lemmatized versions and finally the condensed versions of the texts were compared.

Beyond the comparison of the original texts to their adaptations different translations to the same language are also compared to each other. This kind of comparison might reveal differences in connection with changes in vocabulary. We would be able to find those sections of the texts where the translation is richer or shallower in vocabulary then the original text or another translation. To find explanation for the presence and hiatus of the vocabulary rich segments of the texts the appearance of hapax legomena was also tested. The method for the distribution of hapax legomena were similar to that of the word types.

At this stage of the project the different adaptations of The Jungle Books (Rudyard Kipling), The Da Vinci Code (Dan Brown), The Adventures of Tom Sawyer (Mark Twain), The Adventures of Robinson Crusoe (Daniel Defoe), Alice's Adventures in Wonderland, Through the Looking-Glass and What Alice Found There (Lewis Carroll), Sorstalanság (Kertész Imre) are analyzed, along with some novel without their translations. However, I have to note that the selected texts and the advancing on the texts highly depend on the availability of the printed and the digitized form of them. Most of these texts are manually scanned and digitized because their availability differs due to various reasons. To get comparable results and make the texts readable for DyMoCASAT, processable for the lemmatizer programs the texts should be converted into plain text with well defined borders of paragraphs.

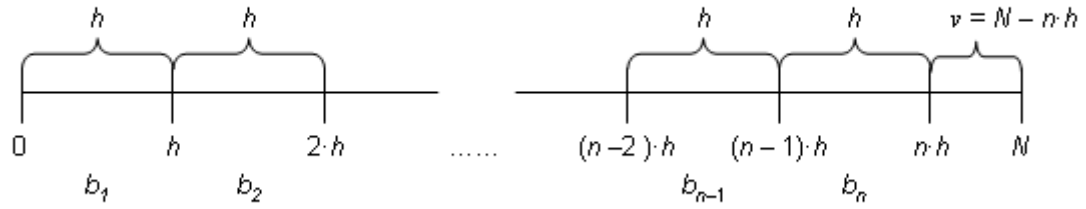

Fig 1. An *N* token-long text is divided into *h*-token-long blocks. The length of the blocks is usually one-hundred tokens. This choice of length is proved to be a good selection because it is longer than the average length of sentences, so syntactic constrains should be taken into account.

III. Results

A. Determining the number of newly introduced word types of a text

To carry out the analysis DyMoCASAT is used [7], [9], [10]. The first step in this process is the analysis of the text. At this stage the number of the tokens (*N*) and then the number of different words (*V*(*N*)) are counted in the text. In the second step the text is divided into intervals of equal length (*h* is the length of an interval, which is usually one-hundred token-long), called blocks.

In each block the different words are identified and the number of their occurrence is counted. The amount of data gained from the text with this method required a time and space consuming data storage. We had to be prepared for further analysis of these data, and had to find a data structure which can be searched with reasonable speed. The solution for this problem is a theoretical three-dimensional matrix. The matrix is theoretical because in practice the data stored in a number of text files. The number of files equals the number of different initial characters found in the text. Each file is identified by these initial characters. This is the first dimension of the matrix. Within each file as many different words are stored as found in the text starting with the same initial character. This is the second dimension of the matrix. Following the paragraph of the word the numbers of occurrence of the word in the blocks are stored. To be able to store numbers greater than ten with only one digits we had to find a numerical system greater than ten. Working with one-hundred token-long blocks even the hexadecimal numerical system seemed small, so we decided on using a system with twenty-seven digits. In this system the numbers are replaced with the letters of the English alphabet and the zero with a character from outside of the alphabet. The last character of the paragraphs indicates the last block where the selected word was found. This is the third dimension of the matrix.

The advantage of storing the occurrence of the words in this theoretical matrix lies again in the numbers. Let us denote the number of different initial characters as k, considering the most frequent initial character as m, and the maximum number of blocks as n. In a real three-dimensional matrix this amount of data requires a $k \times m \times n$ celled storage place. In practice, however, we might have characters which do not start words, the number of paragraphs in each file differs based on the frequency on the initial character, and finally, the length of the paragraphs also differs based on the last occurrence of the word.

With this method all the information about the appearance of the words is stored, so in case the whole text would be restored within the limit of the length of the blocks.

To each block the number of newly introduced words, the number of hapax legomena, the number of different words can be assigned depending on the goal of the analysis.

Our primary aim was to follow the changes of the newly introduced word types, so the number of these words in each block has to be counted (Fig. 2, upper left panel). In general, the number of newly introduced word types follow a monotonic decay. However, we can find sections where this monotonic decay is reversed and a sudden increase then a sudden drop can be detected in the number of the newly introduced words. We are interested in finding the location of these sections and finding reasons for their presence. By mapping the number of newly introduced words, where the domain of the graph is the blocks of the text and the range is the number of the newly introduced words to each block, the protuberances of the graph suggest the location of these unique sessions of the text. Suggestions however, are closer to subjective judgments than objective facts, so we had to find a more reliable analysis of the graphs than just looking at them. The first step following the mapping of the original data is to rule out the accidental rises of the graph. This can be done by the smoothing of the graph (Fig. 2, middle left panel). As the result of the smoothing only the secondary protuberances of the graph are left, which we are interested in. These secondary protuberances can be grouped into two sets. Those which stand for significant changes create the first group, those which not belong to the second. To be able to distinguish the two sets of the protuberances the following method was invented.

IV. Determining the significant changes in newly introduced words

After collecting the data of the selected text using the same program (DyMoCASAT) first-order statistical models can be built to the text. As it was detailed in Mesthods, the assumption of hypergeometric distribution of word provides more reliable data than the polynomial distribution of them. Based on this assumption DyMoCASAT was extended to be

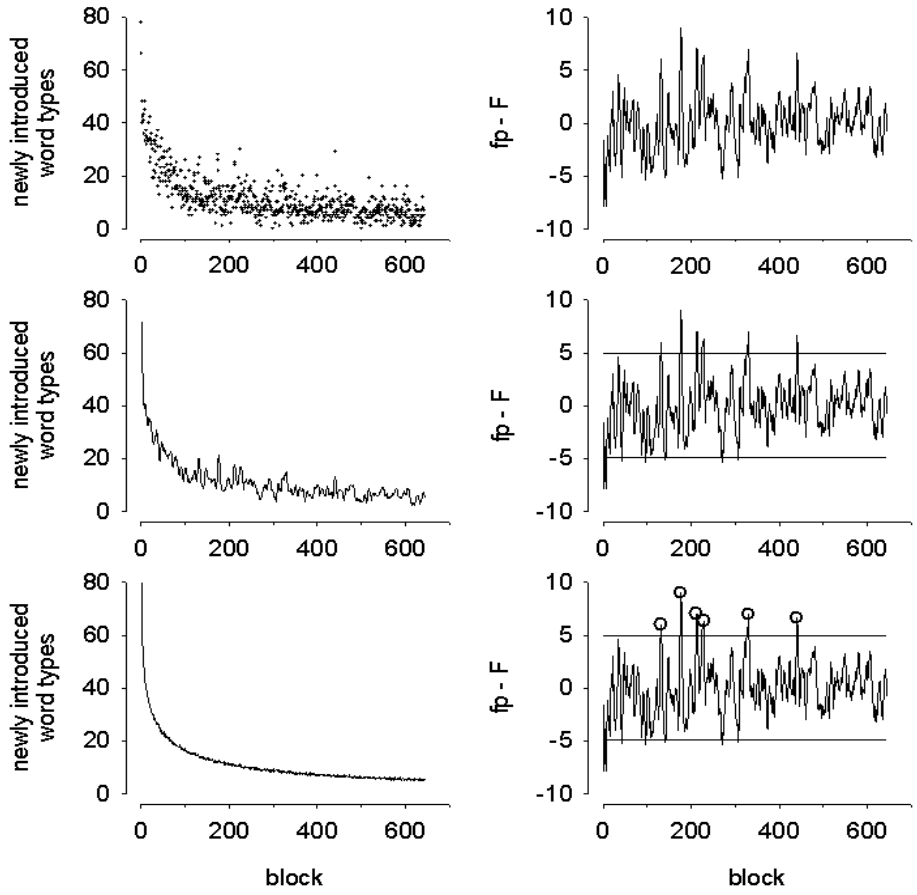

Fig 2. The major steps in the process of analyzing the introduction of word types and lemmas with DyMoCASAT in the condensed English version of The Da Vinci Code. First, the newly introduced words were counted in each block (left, upper), then this function was smoothed (left, middle). In the third step 100 artificial texts were created and the number of newly introduced words was averaged for each corresponding block (left, lower). The difference between the averaged artificial and the smoothed function was calculated (right, upper). The level of significance was determined as M + 2SD (right, middle). Those protuberances were considered significant which exceed the significance level (right, lower).

able to create artificial text. The main characteristic of the artificial text that it carries as many words as the original text with exactly the same frequency. The artificial text is a gibberish, but that is the consequence of the random selection of the words [7]– [9]. To build this model a random selection of words was carried out, the outcome of which was an artificial text.

After creating an artificial text with the same parameters as the original text, the same analysis can be carried out to this text. Using DyMoCASAT, we again are able to create the theoretical three-dimensional matrix, storing all the data considering the appearance of the words. In theory these sets of data are comparable. The random selection, however, by its very nature might cause unpredictable changes in the number of the newly introduced words [27]. To rule this possibility out, a hundred artificial texts were created and averaged ($F(n)$; Fig. 2, left lower graph).

The next step in the process was to determine the differences between the artificial texts and the original ($fp(n)$–$F(n)$; Figure 2 right, upper graph). The mean (M) and the standard deviation (SD) of the differences of $fp(n) - F(n)$ were counted (Fig. 2, right middle graph). The values considered distinguishable are those which exceed the $M + 2SD$ value (Fig. 2, right lower graph). In earlier studies these sudden increases were shown to mark 'longish' inlays in the text and, furthermore, their positions and distribution were found to be unique to the given text [6], [10]).

The result of the analysis is always a graph (Fig. 2, right upper graph – Fig. 2, right lower graph). The flow of this graph is unpredictable, since there is no clear evidence on what it is that makes a writer to increase the number of newly introduced words, and at what point in the story. The corresponding graphs for such pieces of literature always provide protuberances in the graphs.

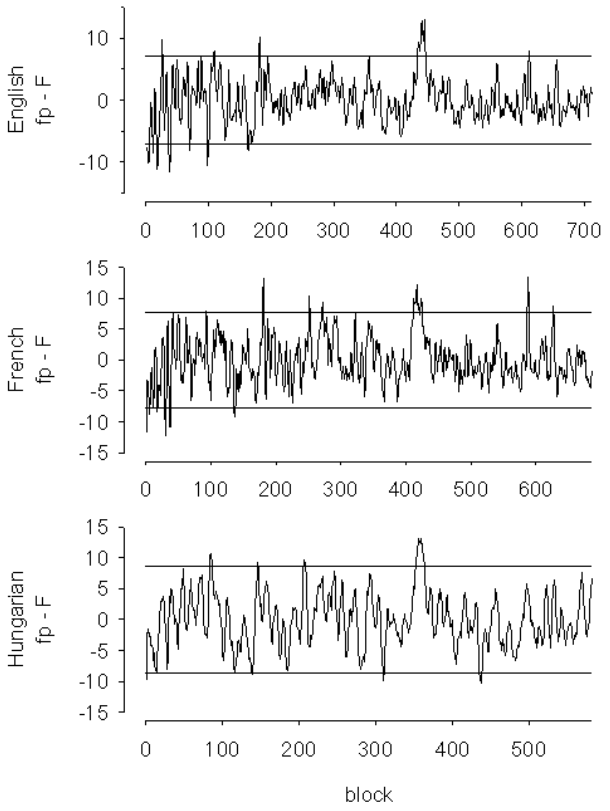

Fig 3. The newly introduced word types of The Adventures of Tom Sawyer and its French and Hungarian translations.

Most of the readers believe that significant changes in the number of newly introduced word types are in close connection with the chapter boundaries. However, in Genette's [16] opinion these vocabulary rich segments of the texts can appear anywhere in the text and carry 'functionally useless' information. This means that if we leave them out the text would still be understandable.

During the proofreading of the texts it was found that these vocabulary rich segments of the text mainly stand for longish descriptions of characters, settings, historical events, and changes in style or language. These vocabulary rich segments of the text are presented on the graph of the newly introduced word types as protuberances. The protuberances have two parameters which are able to describe them. These are the length and the intensity of the protuberances. The length of the protuberances are the number blocks or tokens through which the number of newly introduced word types are over the significance level. The intensity of a protuberance gives the number of newly introduced word types over the significance level. Considering all these we can see that the graph of the newly introduced word types are unique to each text, by them we are able to identify the texts to which they belong. They are like the graphs of sound files. The analyses did not prove that these vocabulary rich segments are expected at the chapter boundaries. They can appear anywhere in the text and the most robust protuberances are proved to be those which are due to changes in style regardless of these boundaries.

In Fig 3 the newly introduced word types of The Adventures of Tom Sawyer are mapped in the original English text, and in the French and Hungarian translations. First the problem of different number of tokens had to be solved ($N_{English} = 711$, $N_{French} = 685$, $N_{Hungarian} = 581$). By the normalization of the domain of the graphs the texts are comparable regardless of their lengths. It is obvious from the graphs that the most robust protuberance is between blocks 434 and 446 in the English text. This protuberance stands for a change in style, the students' writings for their school leaving exam, while the others, smaller both in length and intensively, stand for descriptions. In both translations the change in style is followed remarkably well, while the descriptions are not necessarily. We can find missing protuberances in both languages, and the other way around, protuberances of the translations which were not significant in the original text.

The theory that the protuberances appear at chapter-boundaries was proved wrong by comparing the original texts and their translations where the chapter boundaries of the translated texts are changed as compared to the original text. The vocabulary rich segments of the original and the translated texts stand for the same text segments regardless of the old and new chapter-boundaries.

A. Languages of the texts, word types vs. lemmas

The other question arose in connection with the analysis of the newly introduced word types, and consequently with the vocabulary rich segments of the texts that whether this feature is language independent or not. To test this first DyMoCASAT should be able to read texts of different languages. The primary language of the program is English. However, the program, through its menu, offers the users the opportunity to create their own alphabet, upload this alphabet, and analyze texts in languages different from English. When texts in languages so different in nature are compared the question arose whether the word types are satisfactory enough for such analyses, especially in the agglutinating Hungarian language, due to the fact that affixes attached to the lemmas might increase the number of word types without semantic background. Lemmatization were carried out to English [23], [24] and Hungarian [21], [22] texts. In both languages lemmatizer programs were used to carry out the lemmatization, then the results of the programs were mended to gain comparable data to word types. In this form of the text the lemma and its part of speech tag were concatenated and stored as a single word.

The pattern formed by the newly introduced word types and lemmas show great similarities in both languages [11]. There might be detectable differences between the length and intensity of the vocabulary rich segments. However in general, the vocabulary rich segments of the texts both in the lemmatized and non-lemmatized texts are at the same locations. The only difference found between the English and the Hungarian texts was that at the beginning of the Hungarian non-lemmatized texts the word types might hide significant protuberances by forcing their peaks below the significance level.

B. *Texts and their foreign languages translations*

Analysis of the introduction of new words (either word types or lemmas) in the foreign language translations of a text reveals a novel feature of the translated text. Since texts contain several untranslatable elements, and elements which the translator is not willing to reproduce for various reasons, their replacement with other lexical elements is acceptable [4], [26], [25]. To accept this freedom in the process of translation we can accept the result of the translator program with its serious problems with forming sentences, and finding the suitable words and expressions. From the point of view of translation theory it should be interesting to know how changes in the vocabulary of translations follow the changes of the vocabulary in the original text. With the method described above we can decide whether the translation is as rich in vocabulary as the original text. Our goal is not judging the translation, but checking how faithfully the translation follows the original text in vocabulary. This means that our goal is to highlight those segments of the text which are richer in vocabulary than the main body of the text, and in an already existing translation reveal the shallower and richer text segments.

On the thoroughly analyzed texts (listed in MethodsI) we were able to tell for each translation how faithfully it followed the vocabulary changes of the original texts. On the single texts we were able to find their vocabulary rich segments of them, which might be used for further analyses when their translations are available. However, the found and presented vocabulary rich segments by themselves give useful information for the translators previous to the translation process [12]– [14].

In Fig. 4 the comparison of three different German translations of The Jungle Book to the original text clearly show the differences in the changes of vocabulary. The details of the comparison of these texts reveals the differences of the three translations.

Three different German translations were found for the analysis: Mikush's adaptation from 1951 (from now on Mikush, 1951), and two translations from 1987 from Haef and Harranth (from now on Haefs, 1987, and Harranth, 1987). Among the three texts there is only one which is a full text (Haefs, 1981), unfortunately, the others are cropped to some extent. To be able to make comparison of the three texts they all should have been cropped to the shortest text (Harranth, 1987).

In Fig. 4 the newly introduced word types of the six-story-long English and German translations are mapped. In the order of appearance, the first significant protuberance of the English text stands for the King's Palace in Kaa's Hunting, the second is for the fights of Sea Catch for their territory, the third is the introduction of Rikki-Tikki-Tavi, while the fourth is for the story of Toomai.In Harranth, 1987 all four protuberances are present. However, in this translation new protuberances appeared. One appeared between the original first and second, and stand for the text segment in Tiger-Tiger when Mowgli went to the village. The peak is not too wide, but it is clearly there. The other new protuberance is between the original second and third and represents the

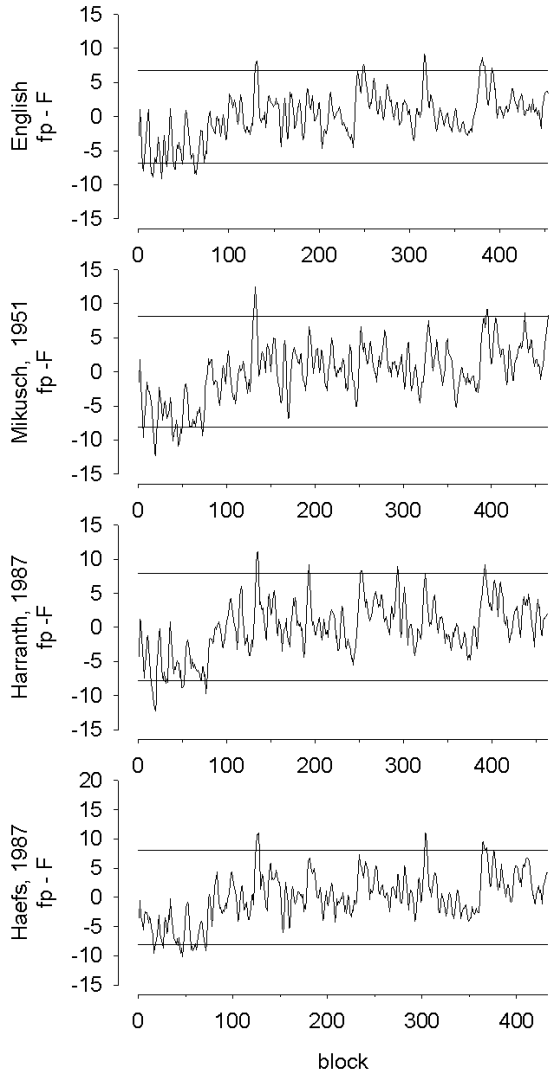

Fig 4 The newly introduced word types of The Jungle Book and its three German translations.

search for the Sea Cow in The White Seal. These two new protuberances mean that we were able to find text segments in this newer German translation which are richer in vocabulary than the original English text. Haefs, 1987 has only three significant protuberances: at the description of King's Palace, Rikki-Tikki-Tavi, and Toomai of the Elephants. However, a closer analysis reveals other remarkable similarities. The missing protuberance from The White Seal is clearly there, but its peak is just below the significance level. Interestingly, the second peak of Harranth, 1987 is also detectable in this version.

Finally, Mikusch, 1951 seems to be the furthest away in vocabulary from the original English text. While the first peak for the King's Palace is clearly there, the next two, although detectable, did not reach the significance level. Finally, the second of the text matches the fourth of the English text. In addition to these, new peaks appeared, which are

unique to this translation: Kala Nag's rush and the song following the story, Shiv and the Grasshopper.

In general, Mikusch, 1951 was found somewhat arbitrary compared to the original English text. The other two translations were able to reproduce the vocabulary rich text segments of the original text much better. Harranth, 1987 goes ever beyond, the translator generated more vocabulary rich text segments than the original text has. Haefs, 1987 seems to be the closest to the original English text.

C. *Texts and their different adaptations*

Beyond the foreign language translations of novels the analyses of other adaptations of the texts might reveal characteristics which should have be taken care of during the adaptations process. In this project the condensed versions of novels were analyzed and compared to the original texts [12], [13]. Being familiar with the vocabulary rich segments of the texts might help the translators on the decisions which segments of the texts should be left out or shortened. These decisions highly depend on the target group. So being familiar with the advance of the vocabulary of a text might be a great advantage compared to subjective decisions.

On the other hand, the method proved to work on the decision in the direct origin of a second level adaptation of a text. This means that by deciding whether the analysis of the newly introduced word types we were able to show the condensed Hungarian translation of The Da Vinci Code originated from the full-length Hungarian translation or the condensed English adaptation. By the analysis of the advance of the vocabulary we were able to show that the condensed Hungarian text is derived from the condensed English text.

D. *The comparison of the results of the analysis with the reviews*

For the German translations of The Jungle Book we were able to find previously published reviews [15], [18]. Considering the vocabulary of the novels are these reviews in accordance with the results of our statistical analysis. These reviews state that the Mikush, 1951 translation carries vocabulary which were not meant in the original text [18]. The closest to the original is Haefs, 1987, while Harranth, 1987 gives a good approximation, but a special interpretation of the text resulting in a vocabulary which gives back the original vocabulary rich segments, but beyond that created new segments which are richer in vocabulary than the original text [15].

This example clearly show that being familiar with the position of the vocabulary rich segments of a the texts in advance to the translation process gives the translator information which segments of the texts requires greater attention because it is further away from the random selection of words than the rest of the text. Gives help in creating condensed versions of the text by finding the 'functionally useless' section of the text.

E. *Hapax legomena in the texts*

It was found that there is a strong correlation between the appearance of the newly introduced word types and the appearance of the hapax legomena, which means in general that if there is rise or a fall in the number of newly introduced word types the same true to the number of hapax legomena. However, rarely there are segments which carry a high number of newly introduced word types and fewer hapax legomena. These words are naturally reused in a later section of the text, and using Genette's expression [16], are less 'useless'. They have function in the text, their specialty is that they were just introduced in that particular block. Those segments which carry both high number of newly introduced word types and hapax legomena are the real 'functionally useless' segments of the text.

V. Summary

The method presented in this article is able to provide those segments of the texts which are richer in vocabulary than the robust part of the text. These segments carry information which is not bound strongly to the flow of the text. They usually give additional information about the settings, the historical background, the characters, or due to severe changes in style. These segments make the novels unique, but if we leave them out the text still would be completely understandable, so they are referred to as 'functionally useless' segments.

To find these vocabulary rich segments of the texts might help us in the analyses of the texts. Beyond that, it was proved that these objective data are able to provide preliminary information for translators by showing the vocabulary rich segments of the texts. Being aware in advance to the translation, or any adaptation of the texts of the position and the intensity of these vocabulary rich segments the translator might pay more attention to them.

VI. References

[1] R. H. Baayen, "The Randomness Assumption in Word Frequency Statistics," In Perissinotto, G. (ed), *Research in Humanities Computing* vol. 5. Oxford: Oxford University Press, 1996, pp. 17–31.

[2] R. H. Baayen, "The Effect of Lexical Specialization on the Growth Curve of the Vocabulary," *Computational Linguistics* vol. 22, pp. 455–480.

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

[4] I. Bart, and K. Klaudy, (ed.) *A fordítás tudománya*. Tankönyvkiadó, Budapest, 1985.

[5] R. Beaugrande, de and W. Dressler, *Introduction to text linguistics*. Bevezetés a szövegnyelvészetbe. Siptár, Péter. (trans.) (2000) Corvina, Budapest, 1981.

[6] M. Csernoch, "Természetes nyelvi szövegek összehasonlítása elsőrendű statisztikai modellekkel," *Publicationes Universitatis Miskolcinensis, Sectio Philosophica, Tomus X. – Fasciculus 3*. Miskolc 2005.

[7] M. Csernoch, "The introduction of word types and lemmas in novels, short stories and their translations," http://www.allc-ach2006.colloques.paris-sorbone.fr/DHs.pdf *Digital Humanities 2006. The First International Conference of the Alliance of Digital Humanities Organisations*. (5–9 July 2006, Paris), 2006.

[8] M. Csernoch, "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, Hungary, 2006, pp. 53–70.

[9] M. Csernoch, "Seasonalities in the Introduction of Word-types in Literary Works," *Publicationes Universitatis Miskolcinensis, Sectio Philosophica, Tomus XI. – Fasciculus 3*. Miskolc 2006–2007, 11–34.

[10] M. Csernoch, "Dinamikusan kezelhető statisztikai modellek irodalmi művek szóalakjainak vizsgálatára," *Alkalmazott Matematikai Lapok* vol. 24 (2007), 2007, pp. 57–77.

[11] M. Csernoch, "Newly introduced word-types and lemmas in Dan Brown's The Da Vinci code and its translations," *Across Languages and Cultures* vol. 8 (2), 2007, pp. 195–220.
[12] M. Csernoch, "Condensed versions of literary works," In *When grammar minds language and literature*. University of Debrecen, 2007d, pp. 107–118.
[13] M. Csernoch, "Újonnan bevezetett szóalakok és lemmák Dan Brown The Da Vinci Code című művében és fordításaiban," *Fordítástudomány*. 10, 2008, pp. 18–41.
[14] M. Csernoch, A novel way for the comparative analysis of adaptations based on vocabulary rich text segments: the assessment of Dan Brown's The Da Vinci Code and its translations. *Digital Humanities 2008*. pp. 95–96. http://www.ekl.oulu.fi/dh2008/Digital%20Humanities%202008%20Book%20of%20Abstracts.pdf
[15] B. Danken, "Kiplings unsterblicher Klassiker." *DIE ZEIT*, 06.11.1987 Nr. 46, 1987.
[16] G. Genette, *Narrative Discours*e. Cornell University Press, Ithaca, New York, 1995, pp. 165.
[17] B. Hajtman, *Bevezetés a matematikai statisztikába*. Akadémiai Kiadó Budapest, 1971.
[18] W. Harranth, "Das Dschungelbuch. Nachwort," Aus dem Englischen von Wolf Harranth (1987). Cecilie Dressler Verlag GmbH & KG, Hamburg, 2004, pp. 217–219.
[19] Gy. Meszéna, and M. Ziermann, *Valószínűség elmélet és matematikai statisztika*. Közgazdasági és Jogi Könyvkiadó, Budapest, 1981.
[20] MorphoWord http://www.morphologic.hu/index.php?option=com_virtuemart&Itemid=320&flypage=shop.flypageTab&page=shop.product_details&product_id=133&lang=en (June 1, 2010)
[21] Cs. Oravecz, and P. Dienes, "Large scale morphosyntactic annotation of the Hungarian National Corpus," In Béla Hollósi and Judit Kiss-Gulyás (eds) *Studies in Linguistics*, vol. VI., Debrecen, 2002, pp. 277–298.
[22] Cs. Oravecz, and P. Dienes, "Efficient Stochastic Part-of-Speech tagging for Hungarian" *In Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas, 2002, pp. 710–717.
[23] P. Rayson, Matrix: "A statistical method and software tool for linguistic analysis through corpus comparison," PhD thesis, Lancaster University, 2003.
[24] P. Rayson, "Wmatrix: a web-based corpus processing environmen,". Computing Department, Lancaster University. http://www.comp.lancs.ac.uk/ucrel/wmatrix/, 2005.
[25] J. Ribycki, "Burrowing into Translation: Character Idiolects in Henryk Sienkiewicz's Trilogy and its Two English Translations," *Conference Abstract, The 16th Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities* Göteborg University, Sweden, 2003.
[26] F. S. Simigné, *A fordítás mint közvetítés*. STÚDIÓ Rendezvények és Nyelvtanfolyamok, Miskolc, 2006.
[27] Gy. Solt, *Valószínűségszámítás*. Műszaki Könyvkiadó, Budapest, 1971.