

Table of contents

1. Introduction	9
1.1. A first-order statistical model	9
1.2. Seasonalities of a literary work	10
1.3. Adaptations.....	10
1.4. Practical usage of the analyses	12
2. Words in a text	15
2.1. Vocabulary size, vocabulary richness, words, word types, newly introduced word types	15
2.1.1. Character	15
2.1.2. Basic terms	16
2.1.3. Number of words.....	16
2.1.4. Number and order of words	19
2.2. Large Numbers of Rare Events	21
2.3. Texts in natural languages vs. first-order statistical models	23
2.3.1. The urn model	24
2.3.2. The re-evaluation of the urn model	24
2.3.3. Number of the different word types	29
2.3.4. The behavior of the newly introduced word types	31
2.4. What is a corpus?	33
2.4.1. Definitions of 'corpus'	33
2.4.2. Corpus and corpora	34
2.4.3. Corpora vs. text archives	35
2.4.4. Corpora of different kinds	35
2.4.5. Selected texts.....	37
2.5. Building 'my corpus': sources of texts in electronic form	38
2.5.1. Texts downloaded from e-libraries.....	38
2.5.2. Digitization of texts.....	38
2.5.3. Sharing digitized texts.....	39
3. Newly introduced words	41
3.1. The essence of DyMoCASAT	41
3.1.1. Counting and storing words	41
3.1.2. Creating artificial texts	47
3.1.3. Determining the significant changes in newly introduced words.....	48
3.2. Smoothing the graph of newly introduced words.....	50
3.3. Word types vs. lemmas.....	51
4. Vocabulary rich segments of novels.....	53
4.1. Texts and their foreign language translations.....	53
4.1.1. List of the analyzed texts.....	53

4.1.2.	Sorstalanság and its translations.....	57
4.1.3.	The Jungle Books and their translations.....	64
4.1.4.	Alice's Adventures in Wonderland and Through the Looking-Glass and What Alice Found There.....	88
4.1.5.	The Da Vinci Code and its translations.....	91
4.1.6.	The Adventures of Tom Sawyer	101
4.2.	Lemmatized texts.....	104
4.2.1.	Lemmatized Sorstalanság and Fateless	104
4.2.2.	Lemmatized Alice stories.....	109
4.2.3.	Lemmatized versions of The Jungle Book	111
4.2.4.	Lemmatized The Da Vinci Code.....	115
4.3.	Condensed versions of literary works	118
4.3.1.	Condensed English The Da Vinci Code.....	118
4.3.2.	A comparison of the English and the Hungarian condensed versions of The Da Vinci Code.....	120
4.3.3.	A comparison of the full-length and the condensed texts of The Scarlet Pimpernel.....	122
4.3.4.	A comparison of the full-length and the condensed texts of The Adventures of Tom Sawyer	125
5.	Hapax legomena and newly introduced words.....	129
5.1.	Hapax legomena in the texts.....	129
5.1.1.	The appearance of hapax legomena	130
5.1.2.	The distribution of hapax legomena.....	134
5.2.	Distribution of hapax legomena	140
5.2.1.	First order statistical models	140
5.2.2.	The comparison of the distribution of hapax legomena in English texts and in their models	142
5.2.3.	Hapax legomena of the Hungarian texts and their models.....	149
6.	Appendix.....	153
6.1.	Frequency lists of The Gold Bug.....	153
6.2.	Editing the lemmatized texts	154
6.3.	Lemmatization of the English texts	155
6.3.1.	Cleaning the English lemmatized files.....	155
6.3.2.	Raw lemmatized English text.....	157
6.3.3.	Procedure to delete the non-token rows from the raw lemmatized English text	159
6.3.4.	Rows of non-tokens cleaned from the lemmatized English text.....	160
6.3.5.	Lemmas with the tags of the word types.....	162
6.3.6.	Importing the lemmatized text into an Excel spreadsheet.....	163
6.3.7.	Deleting the uninformative columns	164

6.3.8.	Changing the order of the columns	165
6.3.9.	Lemmas and their tags separated by Tab	166
6.3.10.	Procedure to retrieve the tags of the lemmas.....	167
6.4.	Lemmatization of the Hungarian texts	168
6.4.1.	Cleaning the Hungarian lemmatized files	168
6.4.2.	Raw lemmatized Hungarian text	170
6.4.3.	Procedure to delete the non-token rows from the raw lemmatized Hungarian text	172
6.4.4.	Lemmatized Hungarian text with the rows of non-tokens removed.....	173
6.4.5.	Lemmas with the tags of the word types	174
6.4.6.	Macro to remove the affixes from the words tagged with \unknown tag.....	175
6.4.7.	Hungarian lemmatized text with the \unknown tags removed.....	176
6.4.8.	Lemmas and their tags in the lemmatized Hungarian texts with some superfluous characters.....	177
6.4.9.	Procedure to remove unnecessary characters from the tags ...	178
6.4.10.	Lemmas with their tags in the Hungarian lemmatized texts ..	179
6.5.	Afterwords of Harranth, 1987	180
6.6.	The process of digitization	182
6.6.1.	Scanning	182
6.6.2.	Typing.....	183
6.6.3.	Texts and tools	183
7.	Sources – novels, short stories	185
8.	References	189
9.	E-sources.....	196