Lexical compactness across genres in works by Karel Čapek

Ján Mačutek[1], Michaela Koščová[1], Radek Čech[2]

[1]Department of Applied Mathematics and Statistics, Comenius University, Bratislava, Slovakia

[2]Department of Czech Language, University of Ostrava, Ostrava, Czech Republic

Abstract

We examine the lexical text compactness in 59 texts by the Czech writer Karel Čapek. The lexical text compactness is defined as a ratio of linked pairs of sentences to all pairs of sentences, where two sentences are considered linked if they contain the same content word. Several related text properties are also investigated. The lexical text compactness does not provide a tool for an automatic text typology; on the other hand, its properties display a standard behaviour (e.g., they depend on the text length). A preliminary analysis of interrelations among several properties of the lexical text compactness is provided.

Key words : text studies, text genres, Czech language, Karel Čapek.

1. Introduction

The lexical text compactness (*LTC* hereafter) is a text index introduced by Mačutek and Wimmer (2014). According to the *LTC* definition, two sentences in a text are considered linked if there is at least one word lemma which is contained in both of the sentences. Only content words (nouns, adjectives, non-modal verbs, adverbials) are taken into account. Denote n the number of sentences in a text and L the number of sentence pairs which are linked in the abovementioned sense. Then *LTC* is defined as the ratio of the number of the linked pairs to the number of all pairs of sentences, i.e.,

$$LTC = \frac{L}{\binom{n}{2}}.$$

Obviously, *LTC* can take values between 0 and 1. If every two sentences contain at least one common content word, the text achieves the maximum level of the "lexical compactness" (i.e., $LTC = 1$); on the other hand, in a "lexically loose" text, with the *LTC* close to 0, would consist of sentences which mostly do not share the same word(s). Mačutek and Wimmer (2014) suggested a statistical test for the difference between two *LTC*s and applied it to two short Slovak journalistic texts.

The links as defined above are inspired by works on so-called text aggregates (sometimes called hrebs in honour of the Czech linguist L. Hřebíček, the "founder" of similar text structures), see e.g. Hřebíček (1997), Ziegler and Altmann (2002) and Altmann (2014). However, text aggregates are usually understood more semantically; e.g., in a text on the history of France, all sentences which contain expressions like Louis XIV, le Roi-Soleil, king,

ruler, he, etc. would all be mutually linked, whereas *LTC* focuses on the lexical level only (which makes an automatic text analysis immediately possible using existing lemmatization computer programs; if one wants to include also on the sematic level, the need of creating a special, relatively complicated software tool arises).

In this paper we aim at further analyses of this text index and its properties.

2. Analyzed texts

We analyzed *LTC*s (and other related numerical text characteristics, see Section 3) computed from 59 Czech texts of six genres (nine fairytales, ten texts from each of the following: journalistic texts, private letters, scientific texts on aesthetics, short stories, travel books), all of them written by Karel Čapek.

Karel Čapek (1890-1938) was one of the most famous Czech writer of the 20th century. Although he is probably best known as a science fiction novelist, he produced works of many genres: novels, short stories, poems, dramas, literary reviews, essays, fairytales, scientific texts, travel books etc. He was also a very influential journalist in Czechoslovakia between the two World Wars,.

The texts were taken from the Karel Čapek online project (see the references). Text length varies from 5 to 314 sentences. By choosing texts written in the same language by the same author we try to eliminate (or at least to reduce as much as possible) influences of the language and of the author; the results thus should depend (mainly) on the genre and/or on the text length.

3. Results

Numerical results obtained can be found in Table 1. Texts are denoted as follows: fairytales F1-F9, journalistic texts J1-10, private letters L1-L10, texts on aesthetics AE1-AE10, short stories S1-S10, travel books T1-T10.

Seven text characteristics were considered:

1) The text length (n) is measured in the number of sentences.
2) The number of links in a text (L) is defined in Section 1.
3) The number of linking words in a text *(NLW*) provides additional information to L; if a pair of sentences is linked by more than one word, *NLW* takes into account all of them.
4) The *LTC*, as defined in Section 1.
5) The link redundancy (*LR*) is computed as *NLW* / L.
6) The link length is the difference between positions of respective sentences in a text (e.g. if the first and the fourth sentence in a text are linked, the link length is 3). Given that in the text consisting of n sentences there are n-k possible links of length k, the respective exploitation rate is equal to the number of observed links of length k divided by n-k. Finally the mean of exploitation rates (*ME*) for lengths 1,2,...,k-1 is evaluated.
7) The *SDE*, which is the standard deviation of the exploitation rates.

	n	*L*	*NLW*	*LTC*	*LR*	*ME*	*SDE*
F1	122	2876	4249	0.195	1.477	0.365	0.091
F2	103	1345	1699	0.128	1.263	0.296	0.163
F3	152	1908	2422	0.083	1.269	0.167	0.106
F4	314	7060	9301	0.072	1.317	0.166	0.104
F5	226	4237	5156	0.083	1.217	0.149	0.071
F6	128	2346	3237	0.144	1.380	0.277	0.113
F7	117	1314	1847	0.097	1.406	0.157	0.085
F8	118	1319	1615	0.096	1.224	0.176	0.079
F9	148	3533	4807	0.162	1.361	0.333	0.077
J1	21	70	122	0.167	1.743	0.249	0.18
J2	31	189	262	0.203	1.386	0.42	0.195
J3	20	91	140	0.239	1.538	0.518	0.176
J4	25	196	311	0.327	1.587	0.664	0.129
J5	12	13	23	0.098	1.769	0.17	0.166
J6	25	135	176	0.225	1.304	0.453	0.212
J7	12	24	34	0.182	1.417	0.409	0.233
J8	5	5	12	0.250	2.400	0.333	0.471
J9	7	9	14	0.214	1.556	0.519	0.293
J10	10	12	19	0.133	1.583	0.223	0.21
L1	22	101	174	0.219	1.723	0.391	0.18
L2	15	28	35	0.133	1.250	0.23	0.152
L3	74	536	668	0.099	1.246	0.19	0.127
L4	27	79	107	0.113	1.354	0.252	0.193
L5	13	3	4	0.019	1.333	0.024	0.12
L6	27	103	124	0.147	1.204	0.301	0.153
L7	16	44	73	0.183	1.659	0.43	0.257
L8	13	34	67	0.218	1.971	0.391	0.25
L9	28	81	126	0.107	1.556	0.175	0.129
L10	11	23	38	0.209	1.652	0.334	0.208
AE1	84	1198	1772	0.172	1.479	0.323	0.126
AE2	63	809	1217	0.207	1.504	0.384	0.143
AE3	66	568	685	0.132	1.206	0.267	0.131
AE4	25	124	160	0.207	1.290	0.337	0.194

AE5	81	738	931	0.114	1.262	0.193	0.105
AE6	54	523	737	0.183	1.409	0.347	0.126
AE7	24	95	157	0.172	1.653	0.253	0.197
AE8	28	127	178	0.168	1.402	0.276	0.165
AE9	22	84	115	0.182	1.369	0.321	0.155
AE10	44	368	509	0.195	1.383	0.349	0.141
S1	132	1356	1849	0.078	1.364	0.156	0.068
S2	112	1492	2040	0.120	1.367	0.254	0.138
S3	123	1724	2830	0.115	1.642	0.219	0.077
S4	159	2453	3416	0.098	1.393	0.204	0.071
S5	133	1745	2156	0.099	1.236	0.194	0.077
S6	82	930	1179	0.140	1.268	0.236	0.105
S7	258	1364	1637	0.021	1.200	0.051	0.057
S8	115	1627	2302	0.124	1.415	0.239	0.083
S9	217	3667	4564	0.078	1.245	0.141	0.078
S10	157	2693	3459	0.110	1.284	0.212	0.081
T1	38	150	176	0.107	1.173	0.175	0.115
T2	27	114	147	0.162	1.289	0.295	0.161
T3	27	122	173	0.174	1.418	0.289	0.156
T4	24	79	100	0.143	1.266	0.253	0.134
T5	27	147	213	0.209	1.449	0.41	0.208
T6	27	65	93	0.093	1.431	0.126	0.117
T7	15	42	89	0.200	2.119	0.429	0.245
T8	26	50	70	0.077	1.400	0.163	0.206
T9	21	80	122	0.190	1.525	0.349	0.172
T10	25	51	71	0.085	1.392	0.139	0.106

Table 1. LTC and related text characteristics.

Cluster analysis was applied to different subsets of the characteristics from Table 1, however, clusters obtained did not offer a reasonable interpretation. There are several possible explanations. First, the influence of the author can be - with respect to the *LTC* and related text properties - much stronger than the influence of the genre (we remind that all texts analyzed were written by the same author). Second, the *LTC* can be a text property with no straightforward interpretation and application, i.e., it can depend on many other factors, not only on the author or genre.

Theoretical implications of the results obtained are much more interesting. The *LTC* and its "relatives" clearly depend on the text length (which is true also for many - or almost all - other text properties, indices, etc., see e.g. Čech, 2015, and references therein).

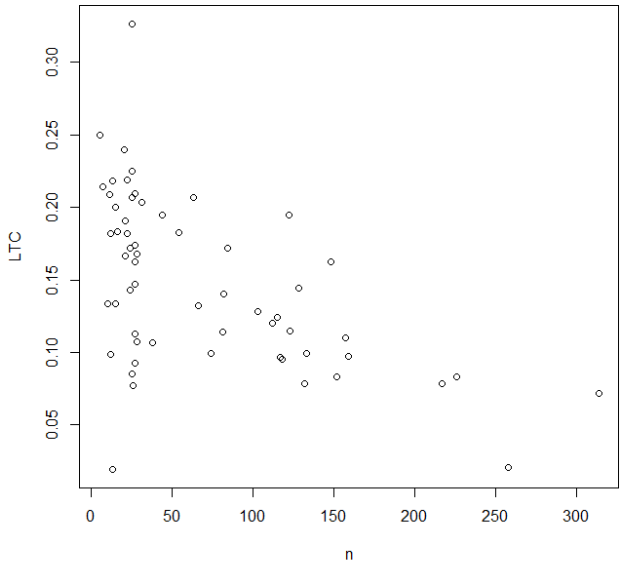

Figure 1. Dependence of LTC on text length

.

Figure 1 shows that the LTC tends to decrease with the increasing text length. Moreover, the same is true for the variability - while short texts display a wide range of *LTC*, values of the index seem to become more stabile for longer texts.

One can see a very similar pattern in Figure 2 (the dependence of the link redundancy on the text length). Also for this text property it holds the longer the text, the smaller its value, and the variability of *LR* is much higher for short texts.

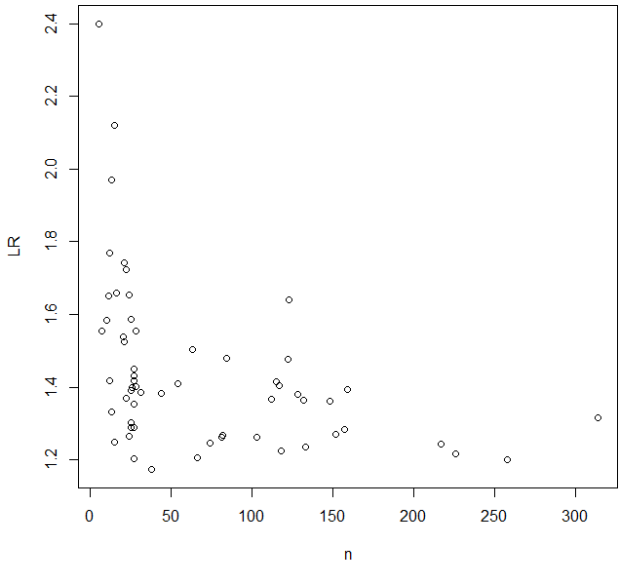

Figure 2. Dependence of LR on text length.

The same tendency can be, again, observed in the relation between the mean exploitation rate (Figure 3).

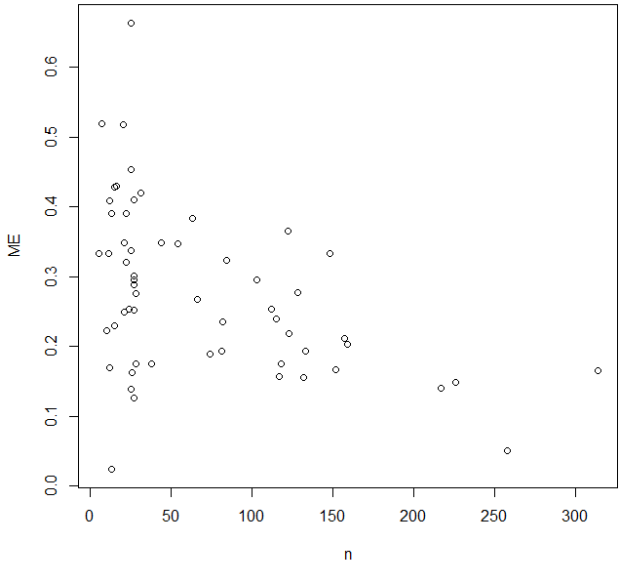

Figure 3. Dependence of ME on text length.

Based on the tendencies displayed in Figures 1 and 3 (both the *LTC* and *ME* decrease with the increasing text length), a positive correlation between the *LTC* and *ME* can be deduced.

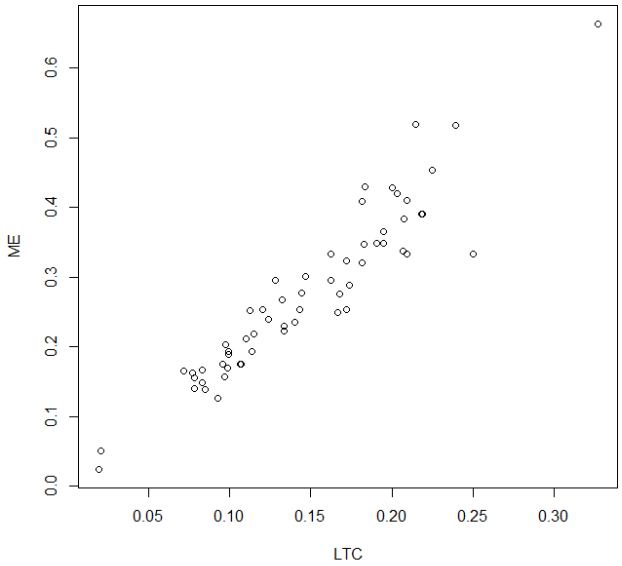

Figure 4. Dependence of ME on LTC.

As can be seen in Figure 4, not only is the correlation positive, but even a linear relation between these two text properties seems to be realistic, at least for the texts used in our study (the Pearson correlation coefficient is 0.94). However, one should remain cautious, as linearity is more an exception than a rule in relations between linguistic units, indices, etc. (see e.g. Mačutek and Rovenchak, 2011, for a discussion why one particular seemingly linear relation is, in fact, non-linear).

There are some text properties which were not considered in this study but which are very likely to have an important impact on the *LTC*.

The first of them is the sentence length measured in the number of words. Quite obviously, longer sentences have a higher probability of establishing new links than shorter ones. Hence, the *LTC* should increase with the increasing sentence length.

Next, also the thematic concentration of a text (see Popescu et al., 2009; Čech et al., 2015). Texts with a high thematic concentration contain relatively few key words (which characterize the topic of a text) which occur with high frequencies, i.e., texts with higher thematic concentrations can be expected to have a higher *LTC*s.

The two hypotheses presented above should, however, be tested on a more diverse language material (i.e., on texts in more languages written by different authors).

4. Conclusion

The *LTC*s in 59 texts written by one authors were examined. We can conclude that neither the *LTC* itself, nor combined with other related text properties is applicable to an automatic text typology.

The paper brings some insights into the relations of the *LTC* to other text characteristics. The text length seems to play a crucial role. In general, the *LTC*, *LR* and *ME* decrease with the

increasing text length. The *ME* seems to depend linearly on the *TLC* (or, if it is a non-linear relation, it is similar to a linear function).

The fact that the variability of the text properties under study displays a high variability for short texts, but for long ones it is more stabile, indicates that the author can have some control over the vocabulary he/she uses to a certain extent; as a text becomes long, some general text laws independent of the author seem to prevail, as the variability decreases. The same behaviour can be observed also for other text properties (see, e.g., Čech, 2015, and references therein).

Hypotheses presented in Section 3 will be tested and mathematical models of the relations mentioned in this paper will be developed when data from more texts are available.

References

Altmann, G. (2014). The study of hrebs. In Altmann, G., Čech, R., Mačutek, J. and Uhlířová, L., editors, *Empirical Approaches to Language and Text Analysis*, pages 1-13. RAM-Verlag.

Čech, R. (2015). Text length and the lambda frequency structure of a text. In Mikros, G.K. and Mačutek, J., editors, *Sequences in Language and Text*, pages 71-87. Mouton de Gruyter.

Čech, R., Garabík, R. and Altmann, G. (2015). Testing the thematic concentration of text. Journal of Quantitative Linguistics 22(3),

Hřebíček, L. (1997). *Lectures on Text Theory*. Oriental Institute of the Academy of Sciences of the Czech Republic.

Karel Čapek on-line. http://www.mlp.cz/cz/projekty/on-line-projekty//karel-capek/ A common project of the Prague City Library, Institute of the Czech National Corpus (Faculty of Arts, Charles University in Prague), Společnost bratří Čapků, and Památník Karla Čapka (accessed on 5 February 2016).

Mačutek, J. and Rovenchak A. (2011). Canonical word forms: Menzerath-Altmann law, phonemic length and syllabic length. In Kelih, E., Levickij, V. and Matskulyak, Y., editors, *Issues in Quantitative Linguistics 2*, pages 136-147. RAM-Verlag.

Mačutek, J. and Wimmer, G. (2014). A measure of lexical text compactness. In Altmann, G., Čech, R., Mačutek, J. and Uhlířová, L., editors, *Empirical Approaches to Language and Text Analysis*, pages 132-139. RAM-Verlag.

Popescu, I-I., Altmann, G., Grzybek, P., Jayaram, B.D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L. and Vidya, M.N. (2009). *Word Frequency Studies*. Moutun de Gruyter.

Ziegler, A. and Altmann, G. (2002). *Denotative Textanalyse*. Praesens.