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 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 semantic 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 writers 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

Seven text characteristics were considered (their numerical values can be sent upon request):

1) The text length is measured in the number of sentences.

2) The number of links in a text is defined in Section 1.

3) The number of linking words (word lemmas) in a text, denoted as *NLW*, provides additional information to *L*; if a pair of sentences is linked by more than one word, the NLW takes into account all of them.

4) The *LTC*, as defined in Section 1.

5) The link redundancy (*LR*) is computed as $LR = NLW / L$.

6) The mean exploitation rate is defined as follows. First, the link length is computed as the difference between positions of two linked sentences in a text (e.g. if the first and the fourth sentence in a text are linked, the link length is 3). In the text consisting of n sentences, there are n-k possible links of length k. The exploitation rate of links of length k is equal to the number of observed links of length k divided by n-k. The same is done for all possible link lengths (i.e., for lengths 1, 2, ..., k-1) in the text. Finally, the mean of these exploitation rates (*ME*) is evaluated.

7) The standard deviation (*SDE*), which is the standard deviation of the above described exploitation rates.

These characteristics were chosen, because, based on a preliminary analysis of several Slovak texts, they seemed to be promising with respect to an automatic genre classification. One of the aims of this paper was to check whether they can discriminate genres in texts written by the same author, which eliminates the impact of the author.

Cluster analysis was used as a mathematical tool, with almost all possibilities included in the statistical software environment R were applied, i.e., both non-hierarchical (k-means, k-medoids) and hierarchical (divisive and agglomerative clustering) methods with different metrics were applied (see Izenman 2008 for an overview of clustering methods). Also, different subsets of the above mentioned characteristics were used out of two reasons – first, in order to reduce the chance that one particular characteristic distorts the overall picture, second, some of the characteristics are highly correlated (see below). However, regardless of the method used and characteristics chosen, clusters obtained did not offer a reasonable interpretation.

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).

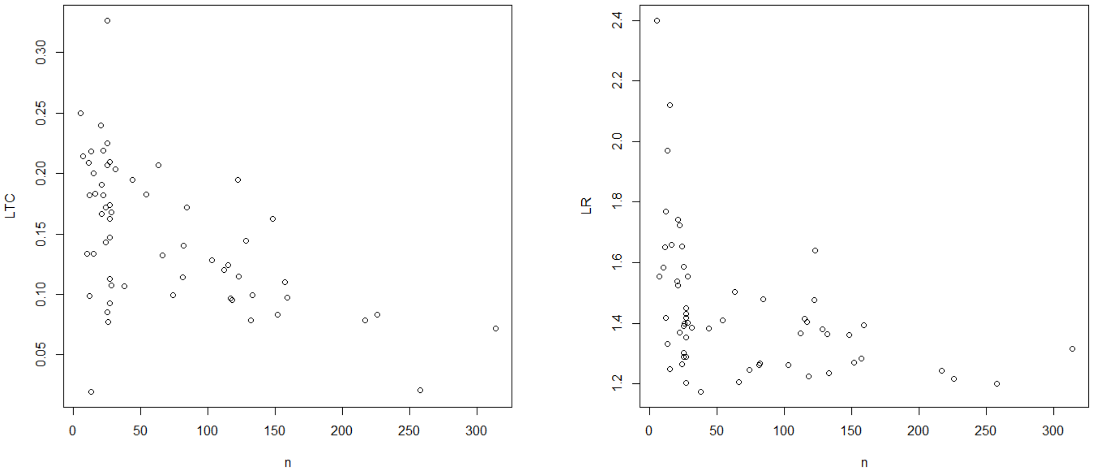

Figure 1. Dependence of LTC (left) and LR (right) on text length.

.

Figure 1 left 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 stable for longer texts. One can see a very similar pattern in Figure 1 right (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.

The same tendency can be, again, observed in the relation between the mean exploitation rate and the text length (Figure 2 left). Based on the tendencies displayed in Figure 1 left and Figure 2 left (both the *LTC* and *ME* decrease with the increasing text length), a positive correlation between the *LTC* and *ME* can be deduced. As can be seen in Figure 2 right, our data confirm this fact.

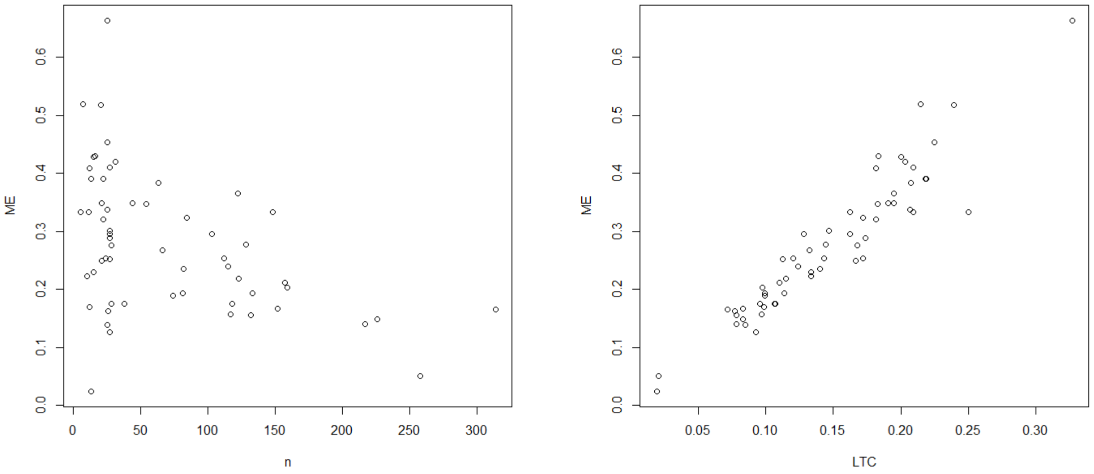

Figure 2. Dependence of ME on text length (left) and on LTC (right).

In addition, 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) can have an influence on the *LTC*. 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 higher *LTC*s.

These two hypotheses 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 author were examined. We can conclude that neither the *LTC* itself, nor combined with other related text properties is applicable to an automatic text typology, at least for the texts under investigation. 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.

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 *LTC* (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 stable, 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 linguistic 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.

Acknowledgement

Supported by VEGA grant 2/0047/15 (J. Mačutek, M. Koščová).

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.

Izenman, A.J. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer.

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.