

Average Word Length from the Diachronic Perspective: The Case of Arabic

Research Article

Jiří Milička

Institute of Comparative Linguistics, Faculty of Arts, Charles University

Received 22 November 2018; Accepted 8 December 2018

Abstract: Previous studies based on English, Russian and Chinese corpora show that the average word length in texts grows steadily across centuries. These findings are in accordance with our results: the average word length in Arabic texts also grows during the analysed time span (8th century to the first half of the 20th century). Our paper shows the detailed statistics of the word length distribution century by century. The dynamics of the average word length correlates with the dynamics of the average word distribution entropy, which encourages an explanation of the phenomenon based on the Shannonian theory of communication.

Keywords: Arabic • word length dynamics • lexicon dynamics • diachronic corpus • entropy • lexical richness

© Sciendo

Introduction

Word length and related features have been studied since the dawn of quantitative linguistics: the word length distribution ([1]: 237, [2]: 438, [3]: 99, [4]), relation between word length and its frequency [5], its relation to the length of its constituent syllables or morphemes [6, 7] or the length of the clause that contains the particular word (see Altmann [8] for the exhaustive bibliography related to this topic).

Therefore, it is peculiar that it has been only recently that the dynamic aspects of the average word length began to be studied from the quantitative linguistic point of view.

The dynamics of the word length from the qualitative point of view is well documented: affixation, the degrading of synsemantic words to endings followed by their merging with autosemantics, and – on the other hand – shortening of endings, borrowing long words from morphologically richer languages or short words from more analytical ones, language ‘purification’ from these borrowings, etc. These and other processes play a role in forming the overall average word length. However, often, several processes co-occur and compete (e.g. ending dropping and Norman borrowings co-occurring in Middle English), making it difficult to determine the size of their effect.

For this reason, it is necessary – and potentially surprising – to measure this feature in a corpus.

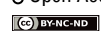
Bochkarev et al. [9] explored the dynamics of average word length in English and Russian texts and concluded that the word length grew steadily through the past 3 centuries. Curiously enough, Chen et al. [10, 11] arrived at the same conclusion for Chinese, although they explored texts over several thousands of years.

These results prompt the search for a generalisation. First, however, the phenomenon needs to be explored in other languages. Hence, we measured the average word length dynamics in a large Arabic diachronic corpus (Corpus Linguae Arabicae Universalis Diachronicus [CLAUDia]). Our results are not surprising considering the aforementioned studies – the average word length increases in Arabic texts as well.

The data obtained for the average word length dynamics across the centuries correlate with the dynamics of the average entropy of the word frequency distribution in this time span. Building upon this finding, we formulate an explanation for the phenomenon based on the Shannonian theory of communication.

* Corresponding author: Jiří Milička, E-mail: jiri@milicka.cz

Open Access. © 2018 Jiří Milička, published by Sciendo.

 This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License.

Data

The diachronic data were obtained from the historical Arabic corpus CLAUDia. CLAUDia contains about 2000 works consisting of approximately 420 million words. It covers a time span from the 7th century to the mid-20th century. However, the 7th century was omitted from this study, for its content is very limited, as well as the *Periodicals Collection* (the genre is strictly limited to the end of 19th and 20th century) and *Thousand and one nights* (it is hard to subsume them to any specific century). Detailed information about the corpus can be found in the study by Zemánek and Milička [12, 13].

Table 1. Numbers of words in the CLAUDia.

Century	Number of words (tokens)
8	2,908,302
9	22,383,232
10	38,487,076
11	38,976,051
12	27,124,049
13	44,606,363
14	56,238,708
15	53,365,460
16	32,854,726
17	17,569,342
18	9,143,613
19	29,344,593
20	19,374,763

There are three main limitations of this data set.

1. The large general diachronic corpora are inherently heterogeneous in terms of style, genres and prevalent topics. The society that uses the language changes and the corpus reflects its change along with the change of language. Consequently, it is very difficult to distinguish between the two. Weighting the corpus according to the genres is very difficult, as the popularity and production of the different genres change over time. Thus, we do not postulate the language–society shift dichotomy and regard the process holistically.
2. The corpus consists of written texts exclusively. We are aware that spoken text features can systematically differ from written text features. Thus, we cannot generalise our findings to the language as a whole.
3. The third limitation is connected with the second one. The Arabic script is (unlike English) well correlated with the spoken text. However, only a small part of the Arabic corpus contains vowel marks. We removed all vowel marks altogether instead of inserting missing vowels. This approach solved the problems with heterogeneity – the percentage of vowel marks in various texts was uneven. The vast majority of the texts in the corpus were originally written and transmitted without vowel marks; the vowel marks in texts were usually added during a later redaction, some of them even automatically (cf. e.g. AlKhalil diacritizer, [14]). Removing the remaining vowel marks also solved the problem with vowel endings, which (according to the normative grammars) should be used in Standard Arabic (al-fuṣḥá), but which is used very scarcely by readers, especially in certain genres that border on popular literature. Also, the main advantage of the Arabic script without vowels is that the spelling is nearly constant through the centuries.

It is very convenient to measure word length in consonants, as it is a compromise between measuring the word length in phonemes and in syllables. It has always been a dilemma whether to use phonemes or syllables [4]. Measuring the number of consonants means that we measure lossily compressed information content of the word. Moreover, the short vowels are only three and their quality can be inferred from the position in the word by a recipient (that is why an Arabic reader is able to interpret Arabic written text without vowels).

There is a strong positive correlation between the word length measured in consonants and the word length measured in phonemes anyway. We have compared 1000 words from an extract of *Kalila and Dimna* with its phonological transcription (the data come from Milička [15], p. 143), and the Pearson's correlation coefficient is $R=0.853$ for individual words.

Results and discussion

Time–word length relation

As can be seen in Fig. 1, the overall trend in the word length dynamics is increasing, which is in accordance with the findings of Chen et al. [11] and Bochkarev et al. [9]. The units of measurement are consonants, as discussed in the previous section.

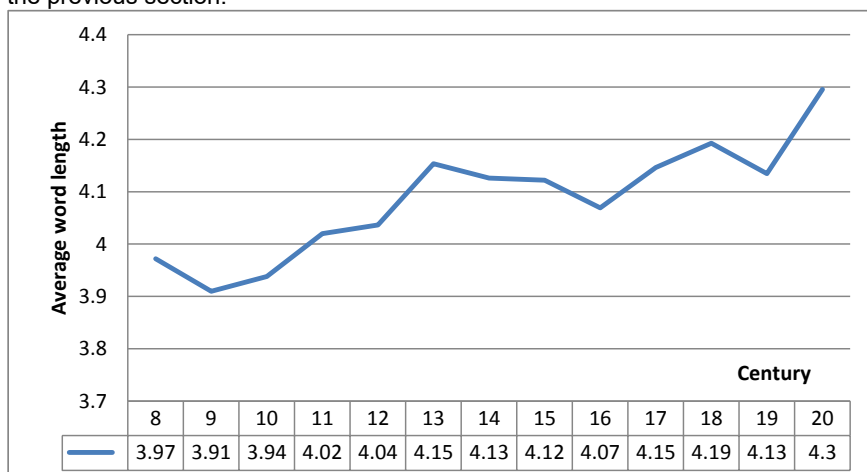


Fig. 1. The dynamics of average word length in the diachronic Arabic corpus.

The increase in the length should mean that the relative frequency of short words decreased (see Fig. 2 as an example) and the relative frequency of long words increased (see Fig. 3 as an example).

Normalisation is necessary in order to show the differences in dynamics between various word lengths in one chart. Therefore, we normalise the relative frequency so that the maximal value is assigned the number 1 and the remaining values are then expressed by a number relative to the maximal value. The relative frequency values for individual centuries (f_c) are transformed into values (f'_c), such as the following:

$$f'_c \doteq \frac{f_c}{\max(f_8, \dots, f_{20})}$$

where c indicates the individual centuries.

This transformation enables us to combine more results into one chart: Fig. 4 depicts the normalised curves for the shorter words (lengths 2–4); Fig. 5 depicts the normalised curves for the longer words (lengths 5–11).

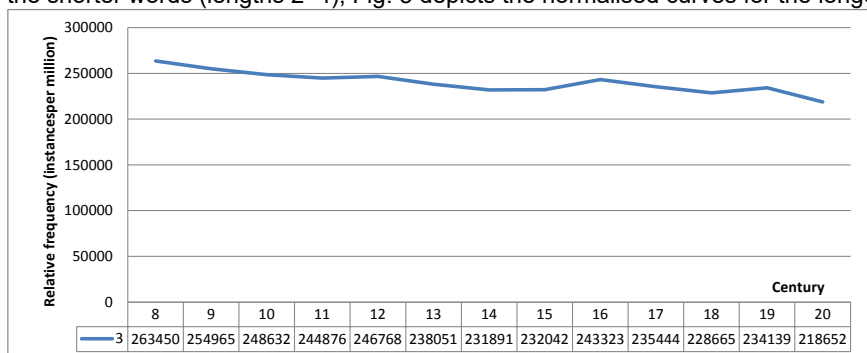


Fig. 2. The dynamics of the relative frequency of three-consonant words in the Arabic diachronic corpus.

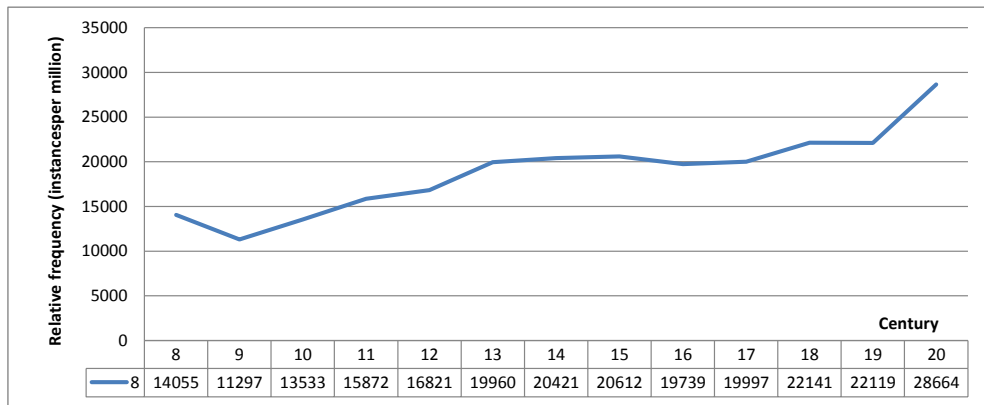


Fig. 3. The dynamics of the relative frequency of eight-consonant words in the Arabic diachronic corpus.

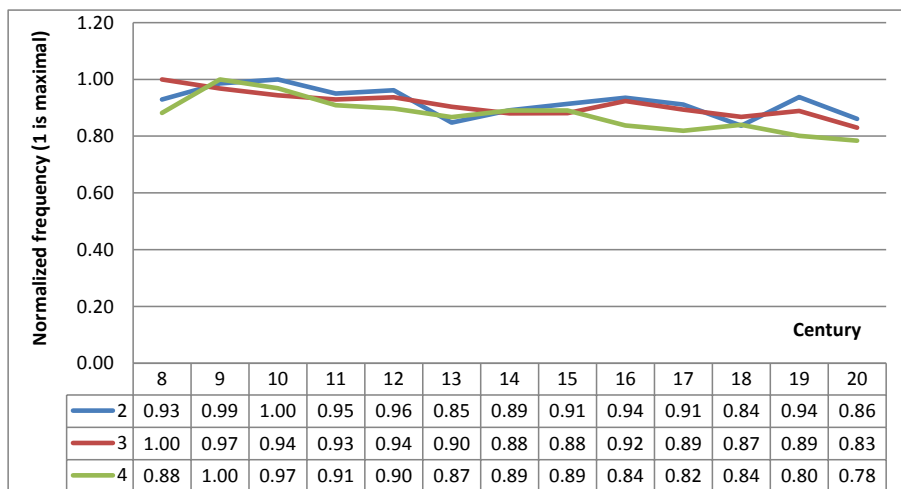


Fig. 4. The dynamics of the relative frequency of short words in the Arabic diachronic corpus.

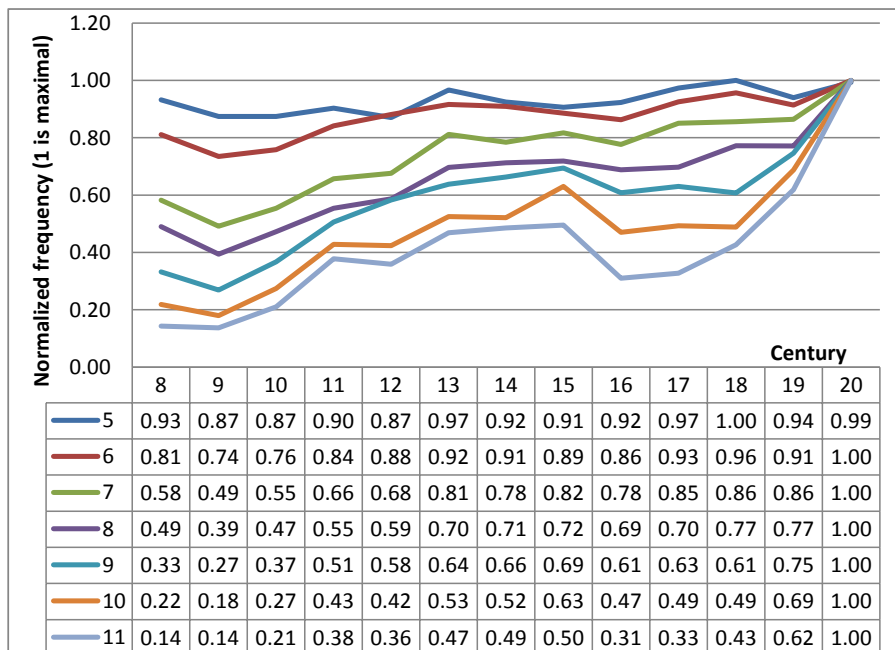


Fig. 5. The dynamics of the relative frequency of long words in the Arabic diachronic corpus.

The same data can be viewed also from the perspective of word length distribution. The distributions for the 8th century can be found in Fig. 6. The data were fitted by the 2-displaced Poisson Distribution model, which is a slightly modified classical Chebanow's model (see Grzybek [4] for an exhaustive theoretical and historical introduction). Usually, the displacement is by only one position (i.e. 1-displaced Poisson Distribution), but as one-consonant words are extremely rare in Arabic, the 2-displaced Poisson Distribution model performs much better for our data. The only role of the model is to give some baseline to which the data can be optically related; therefore, we do not fit the model by the method of least squares or any other standard fitting method. Instead, we derive the only parameter that determines the shape of the Poisson distribution (λ) from the empirical average word length in the given century. We are far from claiming that the 2-displaced Poisson Distribution is an appropriate model for our data; the Coefficient of Determination for the 8th century is $R^2 = 0.988$, while the discrepancy coefficient $C = 0.099$. The fitting of the model for the later centuries is worse, which is probably due to the greater heterogeneity of the corpus in this time span.

The comparison of the data with the model shows that the word length distribution in later centuries has a 'thicker tail' than in the former centuries (see Figs. 7–9). This would suggest that it is the long words that contribute to the increase in average word length. This is in accordance with the data presented in Fig. 5. However, this rather subjective observation deserves further quantification and better analysis, which is beyond the scope of this paper.

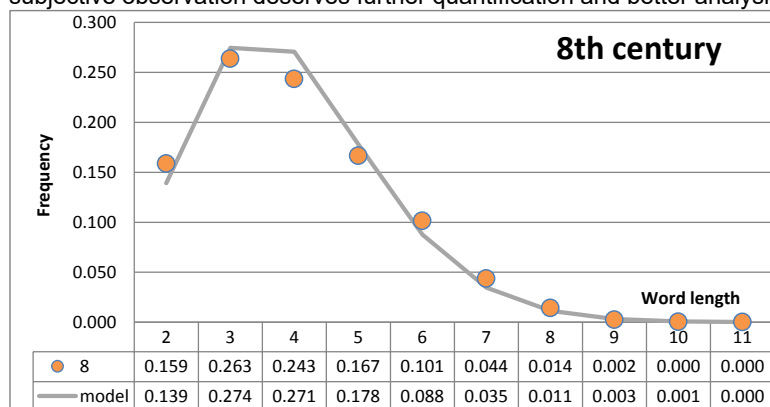


Fig. 6. Word length distribution in the 8th century and its 2-displaced Poisson Model ($\lambda = 1.9717$).

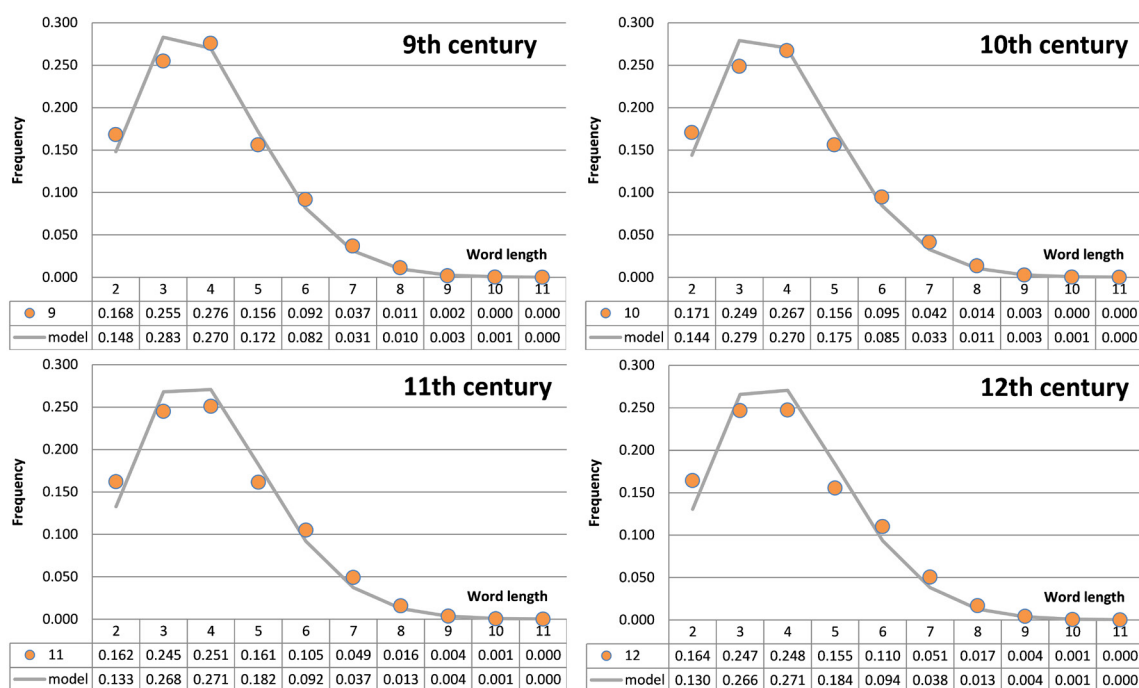


Fig. 7. Word length distributions in the 8th–12th centuries.

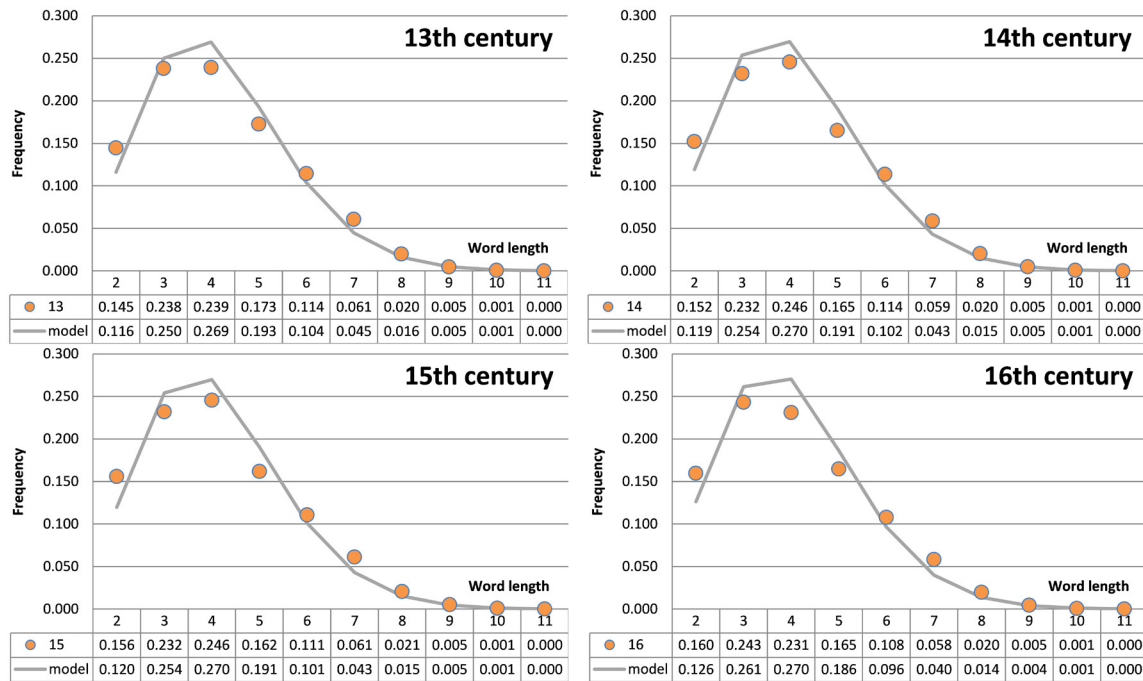


Fig. 8. Word length distributions in the 13th–16th centuries.

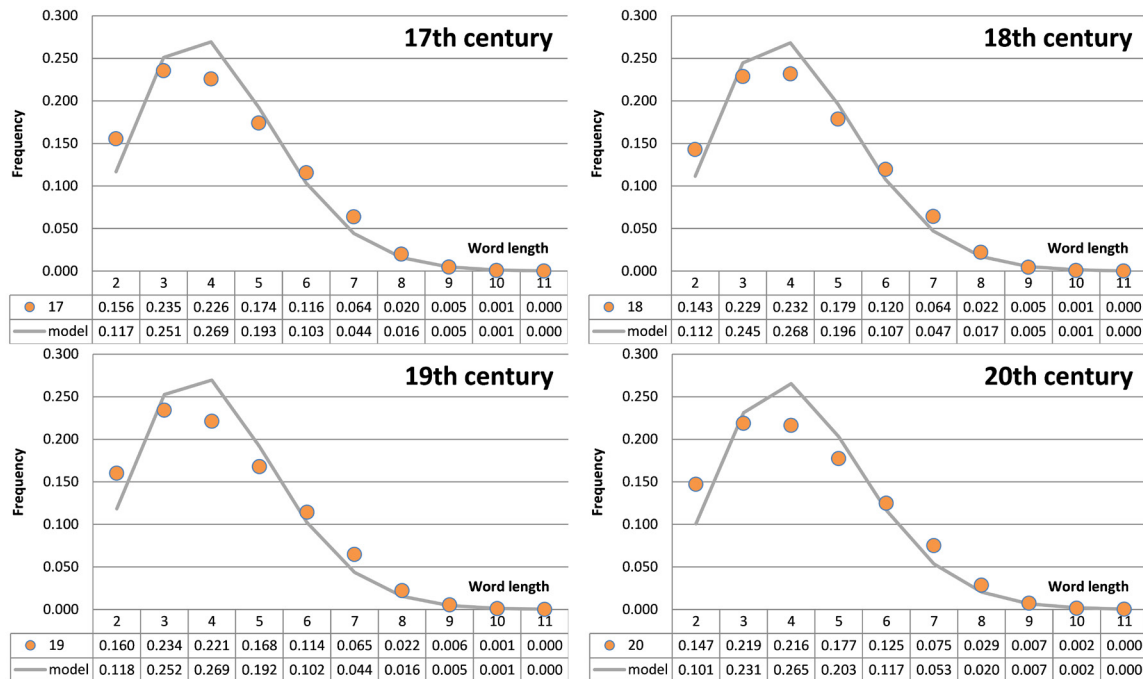


Fig. 9. Word length distributions in the 17th–20th centuries.

Entropy–word length relation

The average word length is closely related to lexical richness. This is in accordance with Zipf's Law of Abbreviations [5] and the study by Kanwal et al. [16]. The following statement illustrates this idea: a greater number of low-frequency words in a text (which means a higher entropy of the word frequency distribution) implies that these low-frequency words are longer than the average words.

In this case, we will explore the mutual relations between the entropy of the word frequency distribution and the average word length. The entropy is used as a metric approximating the least number of bits that can specify the average word in the text; therefore, we can expect that the information needed to specify a word should correlate positively with the amount of information that is actually used to specify the word in real texts (i.e. with word length). More complex vocabulary needs to be encoded by longer words, so that the average relative redundancy remains the same. This relation was studied by Piantadosi et al. [17] (see Ferrer-i-Cancho and del Prado Martín [18] for criticising the approach). Unlike the present study, Piantadosi's study focussed on the individual word length level instead of average length.

The average word frequency distribution entropy was measured in 1000 long chunks of texts to eliminate the influence of text length variability. We have used the MaWaTaTaRaD software [19]; first introduced by Kubát and Milička[20]. The Pearson's correlation coefficient confirms our prediction (the correlation between the two variables being $R = 0.931$), which is illustrated by Fig. 10.

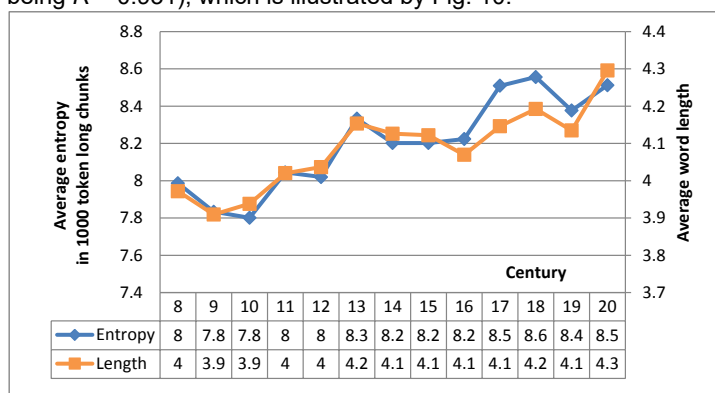


Fig. 10. The dynamics of the average entropy in 1000 token long chunks compared to the average word length (diachronic Arabic corpus).

These results encourage us to explore the correlation between the average word length and the entropy of the word frequency distribution at the level of individual texts. We have taken into account all those texts in the corpus that are longer than 1000 tokens (so that at least one chunk of 1000 tokens can be measured).

Despite the heterogeneity of the text collection (a number of genres, topics and large time spans), the Pearson's correlation coefficient is still convincing ($R = 0.753$), albeit smaller than in the previous case. Fig. 11 depicts the dependency of the average word length on the average entropy of word distribution – each dot in the scatter plot represents one text. The heterogeneity is fairly visible in the chart: there are two attractors of average word length for texts with a higher average entropy of word distribution.

We can identify the texts that have a high average word distribution entropy but the average word length is relatively low: these texts mostly belong to the genre of '*adab*' *belles-lettres* (see Zemánek and Milička [12] for an explanation) and related genres: collections of proverbs, entertainment, morals, education of princes (mirrors of princes) and travels (see Fig. 12). Curiously enough, these genres cluster together with linguistic literature and lexicons, while other scientific fields, along with history, biographies and treatises about culture, cluster together with texts about religion, which form the majority of the corpus.

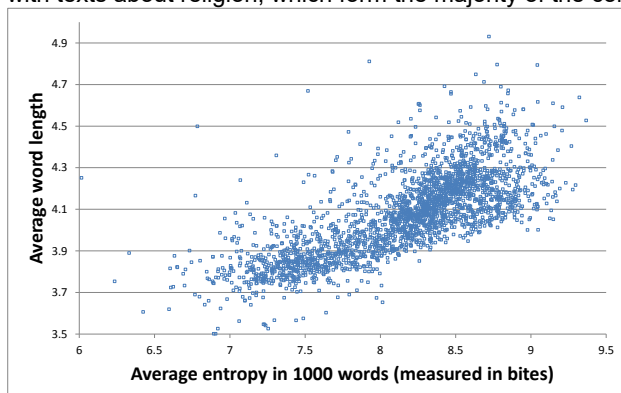


Fig. 11. Dependency of the average word length on the average entropy of word distribution. Each data point represents one text longer than the 1000 word token.

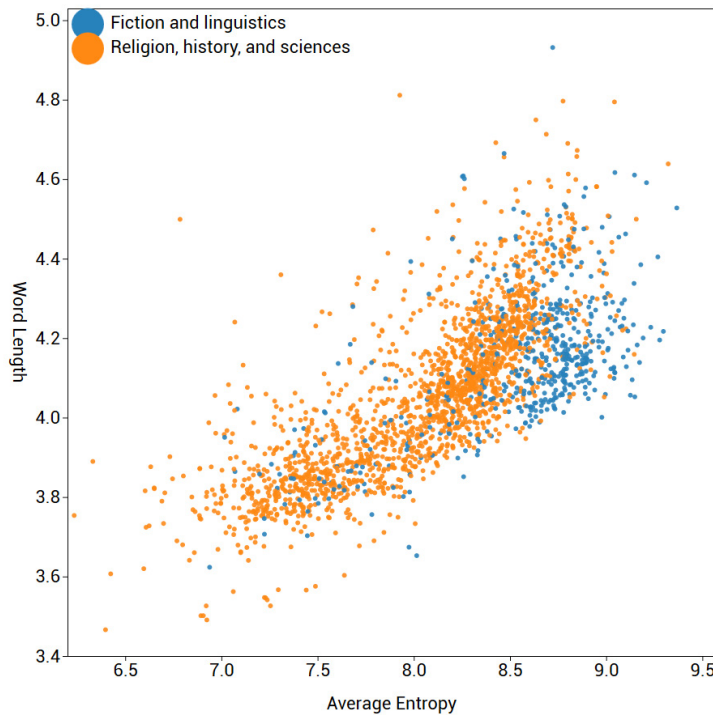


Fig. 12. Dependency of the average word length on the average entropy of word distribution. The genre groups are colour-coded.

Conclusions

We are far from concluding that the increase of the average word length is a general law of the word length dynamics. But we can expect this pattern in other corpora of various other languages, at least for the past centuries. This is due to the increase in the complexity of society, along with the complexity of languages, which are part of the respective societies' cultures. Language complexity manifests itself in lexical complexity, i.e. the entropy of the word frequency. In accordance with the Shannonian theory of communication, the word frequency distribution entropy is strongly positively correlated with the average word length.

It is possible that the word length peak in the 13th century, which we see in our data (Fig. 1), is only an artefact of a bad representativeness of our corpus, but we can find a good explanation for the decrease during the 14th, 15th and 16th centuries – the Mongolian invasion caused the decline of societies that used Arabic as a literary language (i.e. not only Arabs, but also Persians etc., who suffered an immense human and ecological catastrophe).

Remarkably, the average word length and average entropy of the word distribution, despite a strong mutual correlation of these variables, are able to cluster texts according to their genres. This under-examined feature deserves the attention of both linguists and natural language processing (NLP) engineers.

Acknowledgements

The research reflected in this article was supported by the GAČR (Czech Science Foundation; project no. 13-28220S). I also thank Petr Zemánek for various important comments and Denisa Šebestová for proofreading and valuable remarks.

References

- [1] Mendenhall, T.C., 1887. The characteristic curves of composition. *Science*, 9(214), 237–249.
- [2] Elderton, W.P., 1949. A few statistics on the length of English words. *Journal of the Royal Statistical Society. Series A (General)*, 112(4), 436–445.
- [3] Chebanow, S.G., 1947. On conformity of language structures within the Indo-European family to Poisson's law. *Comptes rendus de l'Academie de science de l'URSS*, 55(2), 99–102.
- [4] Grzybek, P., 2007. History and methodology of word length studies. In Grzybek, P. (Ed.), *Contributions to the Science of Text and Language*. Dordrecht: Springer, pp. 15–90.
- [5] Zipf, G.K., 1935. *The Psycho-Biology of Language* (Vol. ix). Oxford, England: Houghton Mifflin.
- [6] Menzerath, P., 1928. Über einige phonetische Probleme. In *Actes du premier Congrès international de linguistes*. Leiden: Sijthoff.
- [7] Altmann, G., 1980. Prolegomena to Menzerath's law. *Glottometrika*, 2(2), 1–10.
- [8] Altmann, G., 2014. Bibliography: Menzerath's law. *Glottology*, 5(2), 121–123.
- [9] Bochkarev, V.V., Shevlyakova, A.V., Solov'yev, V.D., 2015. The average word length dynamics as an indicator of cultural changes in society. *Social Evolution & History*, 14(2), 153–175.
- [10] Chen, H., Liu, H., 2014. A diachronic study of Chinese word length distribution. *Glottometrics*, 29, 81–94.
- [11] Chen, H., Liang, J., Liu, H., 2015. How does word length evolve in written Chinese? *PLoS One*, 10(9), e0138567.
- [12] Zemánek, P., Milička, J., 2014. Quotations, relevance and time depth: medieval Arabic literature in grids and networks. In: *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pp. 17–24.
- [13] Zemánek, P., Milička, J., 2017. *Words Lost and Found. The Diachronic Dynamics of the Arabic Lexicon*. Lüdenscheid: RAM-Verlag.
- [14] Chennoufi, A., Mazroui, A., 2016. Impact of morphological analysis and a large training corpus on the performances of Arabic diacritization. *International Journal of Speech Technology*, 19(2), 269–280.
- [15] Milička, J., 2015. *Teorie komunikace jakožto explanatorní princip přirozené víceúrovňové segmentace textů [The Theory of Communication as an Explanatory Principle for the Natural Multilevel Text Segmentation]*. PhD thesis, Charles University, Prague, Czech Republic.
- [16] Kanwal, J., Smith, K., Culbertson, J., et al., 2017. Zipf's law of abbreviation and the principle of least effort: language users optimise a miniature lexicon for efficient communication. *Cognition*, 165, 45–52.
- [17] Piantadosi, S.T., Tily, H., Gibson, E., 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9), 3526–3529.
- [18] Ferrer-i-Cancho, R., del Prado Martín, F.M., 2011. Information content versus word length in random typing. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(12), L12002.
- [19] Milička, J., 2013. MaWaTaTaRaD, software, available at: <www.milicka.cz/en/mawatatarad/>.
- [20] Kubát, M., Milička, J., 2013. Vocabulary richness measure in genres. *Journal of Quantitative Linguistics*, 20(4), 339–349.