sciendo
DOI 10.2478/jazcas-2021-0059

CZECH TRANSLATIONS OF THE GOSPEL OF MATTHEW FROM THE DIACHRONIC POINT OF VIEW – PLUS ÇA CHANGE…

RADEK ČECH[1] – JÁN MAČUTEK[2,3]
– PAVEL KOSEK[4]
[1] Department of Czech Language, University of Ostrava, Ostrava, Czech Republic
[2] Department of Mathematics, Constantine the Philosopher University, Nitra, Slovakia
[3] Mathematical Institute, Slovak Academy of Sciences, Bratislava, Slovakia
[4] Masaryk University, Brno, Czech Republic

ČECH, Radek – MAČUTEK, Ján – KOSEK, Pavel: Czech translations of the Gospel of Matthew from the diachronic point of view – Plus ça change… Journal of Linguistics, 2021, Vol. 72, No 2, pp. 656 – 666.

Abstract: The paper focuses on dynamics of changes of several linguistic and text properties in diachronic development of Czech. Specifically, we analyze the proportion of identical word-forms (types), the average type length, text length, the proportion of hapax legomena, the moving average type-token ratio, and entropy. For the analysis, seven translations of the Gospel of Matthew from the 14th to the 21st century were used. The study reveals some differences in dynamics of changes of particular properties.

Keywords: Gospel of Matthew, diachronic development, dynamics of changes, properties of text, Czech language

1 INTRODUCTION

Language is a phenomenon undergoing continuous changes. It is well known that there are periods of more intense changes as well as periods of a relative stability of the language system. Further, in particular linguistic planes, one can find differences in the dynamics of changes. For instance, diachronic development of a lexical plane seems to be more or less a continual process, where more dramatic changes, if they appear at all, are usually caused by extra-linguistic factors (such as the National Revival in the 19th century in Czech). On the other hand, in diachronic development of both phonological and morphological planes, we can observe periods of very intense changes followed by periods in which these properties of language remain almost constant.

Up to now, historical linguistics has brought plenty of analyses describing and explaining the development of phonological, grammatical, and lexical properties of language. Following this direction of linguistic research, in this study, we focus on the historical development of some lexical properties and text characteristics which, so far, have not been in the center of attention of historical linguists, at least for

Czech. Our aim is to observe, compare, and explain the dynamics of changes in these properties. For the analysis, we use seven different translations of a single text, namely, of the Gospel of Matthew, which were published from the 14th to the 21st century. This approach allows us to control to a certain extent the impact of factors, such as genre, topic, authorship etc. on the observed phenomena. If we use texts of the same genre, topic etc. (in our case different variants of a single text), we know that potential differences in results are not caused by these factors. For instance, it is well known that the average word length varies, depending on the genre and topic. So, if we want to observe potential changes in word length during historical development of a language, it is difficult to find samples of texts which share the same characteristics in particular periods. Thus, the choice of different variants of a single text seems to be one of acceptable approaches, although of course not the ideal one, cf. [1].

As we mentioned above, dynamics of different language (or text) properties varies. For illustration, let us start with a comparison of the verse 2.1 from the oldest (1) and the youngest (2) Czech translation of the gospel.

(1) *Protož když* ***sě*** *jest* ***urodil*** *Ježíš v* ***Bethlémi*** *židovském za* ***dnóv*** *krále Heroda, tehdy mudráci* ***ode vzchoda*** *sluncě přišli sú do Jeruzaléma.*
(*Bible of Dresden*, 1370s)

(2) *Když* ***se*** *Ježíš* ***narodil*** *v judském* ***Betlémě*** *za* ***dnů*** *krále Heroda, hle, mágové* ***od východu*** *přišli do Jeruzaléma.*
(*Czech Study Translation*, 2009)

‘Now when Jesus was born in Bethlehem of Judaea in the days of Herod the king, behold, there came wise men from the east to Jerusalem.’[1]

At first sight, differences are evident. First, the length of the verse (measured in the number of tokens) differs – 22 tokens in (1) versus 18 tokens in (2). Second, the words that occur in both texts but differ in their form (e.g., *urodil* – *narodil* ‘was born’) are highlighted in bold. Third, there are only nine identical word forms which appear in both (1) and (2), namely: *když* ‘when’, *Ježíš* ‘Jesus’, *v* ‘in’, *za* ‘during’, *krále* ‘king’, *Heroda* ‘Herod’, *přišli* ‘came’, *do* ‘into’, *Jeruzaléma* ‘Jerusalem’. However, there are some properties of language and text which are not so “visible” at first glance, for instance, word length or lexical diversity. Again, for illustration, the average type length in (1) is 4.72 characters, while in (2) it is 5 characters. As for lexical diversity (or vocabulary richness), if we compare the type-token ratio (*TTR*) of the second chapter of the gospel in the oldest and youngest (2) Czech translation,

[1] This is the modern translation of the excerpt, taken from [2].

we get the values of 0.542 and 0.591, respectively. These differences seem to be rather small, especially if one compares them with the number of lexical or word-form changes (see above). So, the question is, whether they follow the variability we observed above or whether they represent more stable properties.

In this study, we start with a comparison of the vocabulary of particular text variants where the differences are most obvious. Then, we focus on both word characteristics (length of word types) and some properties of text (text length and several measures of lexical diversity). As for the analysis of word-type length, our goal is to find out if the relatively high number of phonological, morphological, and lexical changes which have taken place in Czech since the 14th century is followed by changes of this word property. If so, our aim is to observe its dynamics and to compare it with the dynamics of changes in vocabulary. Text characteristics (as opposed to characteristics of language) and their potential changes can be seen from a somewhat broader perspective. They can be explained as a result of a combination of changes in the language system as well as of different translation techniques.

2 LANGUAGE MATERIAL AND METHODOLOGY

For the analysis, we use translations of the Gospel of Mathew from the 14th to the 21st century which were part of

- Bible drážďanská (BiblDrážď), Bible of Dresden (the 70s of the 14th century),
- Bible olomoucká (BiblOl), Bible of Olomouc (1417),
- Bible Melantrichova (BilMel), Melantrich's Bible (1570),
- Bible kralická (BiblKral), Bible of Kralice (1596),
- Bible svatováclavská (BiblSvat), Bible of St. Wenceslas (1677),
- Nový zákon (BiblSýk), New Testament (1909),
- Český studijní překlad Bible (ČSP – Bible), Czech Study Translation (2009).

From the great number of translations of this gospel, we tried to select the ones which are considered significant in terms of development of the Czech biblical translation and which well represent the individual periods ([3], [4]). We decided not to use the first chapter of the Gospel of Mathew because of its specific nature. It introduces the genealogy of Jesus and it mainly consists of the list of names.

The following properties were used for comparison of the texts: 1) the proportion of identical word-forms (types) in pairs of texts (*PIV*), 2) the average type length (*AVL*), 3) text length (*N*), 4) the proportion of hapax legomena (*PHL*), 5) the moving average type-token ratio (*MATTR*), 5) entropy (*H*).

PIV is determined by the number of identical word-forms which occur in two texts. It is calculated as

$$PIV = \frac{|V_i \cap V_j|}{|V_i|},$$

where V is the list of word-forms of the text (i.e., list of types) and i and j are texts. In this study, we calculate *PIV* in relation to the BiblDrážď only (i.e, in all calculations V_i is the set of word-forms which occur in the Gospel of Matthes from BiblDrazd). It allows us to interpret *PIV* as the proportion of changes relative to the oldest translation which is considered as a "reference point" here.

AVL is defined as the arithmetic mean of type lengths in a text. The type length is calculated as the number of characters (e.g., the word *východu* is 7 characters long).

N is determined by the total count of all tokens in a text. A token is a graphic word, i.e., a sequence of characters separated by spaces.

The last three indices are associated with the lexical diversity of the text. Since it is a relatively complex phenomenon which can be seen (and measured) from several points of view, we decided to apply the following methods. *PHL* is the proportion of hapax legomena in the text. *MATTR* is based on the segmentation of text into overlapping chunks (so-called windows), where the type-token ratio is calculated for each window. It is determined as the arithmetic mean of the type-token ratios in all windows, i.e.,

$$MATTR = \frac{\sum_{i=1}^{N-L} V_i}{L(N-L+1)},$$

where N is the number of tokens in the text, L is the length of the window, V_i is the number of types in the window. In this study we use $L = 100$. H is usually interpreted as a measure of system uncertainty. In the case of a text, it expresses the degree of lexical and word form diversity. Specifically, the greater the value of entropy, the more diversified (i.e., less concentrated) the vocabulary is. It is calculated as

$$H = \log_2 N - \frac{1}{N} \sum_{r=1}^{V} f_r \log_2 f_r,$$

where N is the frequency of tokens in the text and f_r is the frequency of a given token.

For the text processing and computation, we used the QuitaUp [5] (for *N*, *PHL*, *MATTR*, *H*) and R-software [6] (for *PIV*, *AVL*).

3 RESULTS

Let us start with the comparison of *PIV,* which should be the most distinct indicator of changes (cf. Section 1). In Table 1 and Figure 1, we can see the

proportions of word forms (types) which BiblOl, BiblMel, BiblKral. BiblSvat, BiblSýk, and ČSP share with BiblDrážď. As we mention in Section 2, we consider BiblDrážď the reference point since it is the oldest surviving Czech translation. The higher the value of *PIV*, the more similar the texts.

Bible	***IWF***	***PIV***
BiblOl	2,299	0.530
BiblMel	1,525	0.351
BiblKral	1,464	0.337
BiblSvat	1,467	0.338
BiblSýk	1,375	0.317
ČSP	1,116	0.257

Tab. 1. Proportions of word forms (types) (*PIV*) which are identical in given texts and the BiblDrážď (which contains 4340 types). *IWF* is a total number of word forms occurring in both BiblDrážď and the given translation

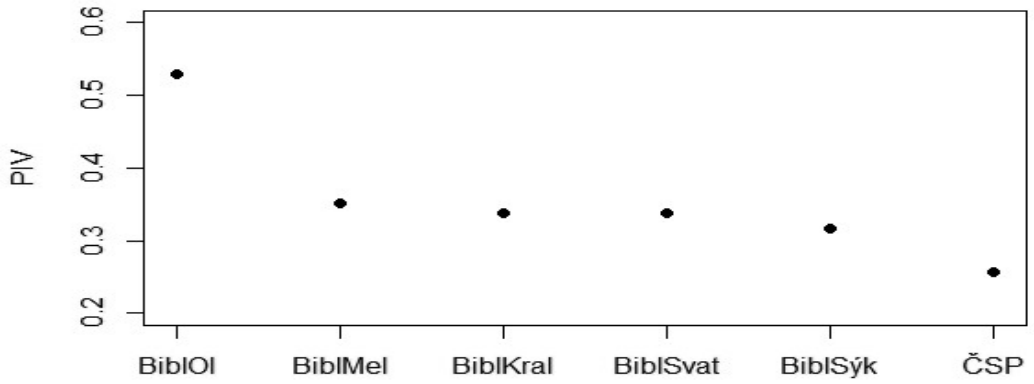

Fig. 1. Proportion of identical word forms (types) (*PIV*) in given texts and the BiblDrážď

The results show that the number of differences is rapidly increasing (i.e., the value of *PVI* is decreasing) from BiblDrážď to BiblMel. The difference between BiblDrážď and BiblOl corresponds with [3, pp. 47, 53] who claim that these two translations were created independently of each other.[2] As for BiblMel, it is considered a translation independent from the older ones. Moreover, there is a 200-year time span between BiblMel and BiblDrážď and more than 150 years between BiblMel and BiblOl, which also can be an important factor. Last but not least, the fundamental phonological, morphological, and syntactic changes which took place in Czech in this period (i.e., from the 14th to the 16th century) influenced the form of BiblMel substantially. By contrast, the next translations up to the beginning of the 20th century

[2] BiblDrážď and BiblOl are copies of the same (lost) source text of the original text of the Old Czech Bible of the 1st edition from the 60s of the 14th century. However, BiblOl differs in the translation of the Gospel of Matthew known as Gospel of Matthew with homilies (from the 70s of the 14th century), cf. [3, pp. 47, 53].

are very similar (in the sense of the method we used). This finding is in accordance with the textual relations between these three translations that are known from secondary sources [3, pp. 180, 212]. The translators/editors of the Gospel of Matthew from BiblKral were influenced by the version published in BiblMel. The translators/editors of BiblSvat were influenced by both BiblMel and BiblKral. BiblSvat, respected by the Catholic Bible translators/editors, was the dominant Catholic translation until the 20th century [4, p. 1887]. All these translations can be interpreted as a result of a specific translation tradition. The "power" of this translation tradition is particularly evident in the fact that each of the BiblMel, BiblKral, and BiblSvat comes from different Czech churches (the Utraquist, Unity of the Brethren, Catholic, respectively). Another factor that has an impact on the language form of the translations is the nature of written Czech. Specifically, a minimum of fundamental changes appear at the phonological and morphological level from the end of the 16th century to the 19th century. Finally, the youngest translation is less similar in comparison to the oldest one. This finding is in accordance with the declared effort of the translators to escape the above mentioned strong translation tradition and to use the language that is commonly used at the beginning of the 21st century. To, sum up, the dynamics of *PIV* development corresponds with current knowledge regarding both the language development and the translation tradition.

Next, let us concentrate on the development of *AVL* (see Table 2 and Figure 2). In comparison to *PIV*, we get a completely different picture. Surprisingly enough, not linguistically important changes of *AVL* took place in the period of more than 600 years. The differences are strikingly small – the biggest difference is the difference of two-tenths of a letter, on average. To get a more complex view of the nature of the data, standard deviations (*SD*) of *AVL* were computed, too. In our case, *SD* expresses a variability of word lengths in particular texts. Again, a great stability of values of *SD* appears throughout the period under investigation. These findings are very unexpected, especially if one realizes both a number of linguistic changes which took place since the end of the 13th century in Czech and diversity of translation strategies. In our analysis, *AVL* seems to be an extremely stable property that resists any development.

Bible	***AVL***	***SD* (*AVL*)**	***N***
BiblDrážď	6.41	2.1	17,327
BiblOl	6.34	2.1	17,818
BiblMel	6.48	2.14	17,548
BiblKral	6.52	2.19	17,621
BiblSvat	6.48	2.15	17,083
BiblSýk	6.52	2.18	17,093
ČSP	6.52	2.16	17,109

Tab. 2. The average type length (*AVL*), the standard deviation of *AVL* (*SD*), and the length (*N*) of individual translations

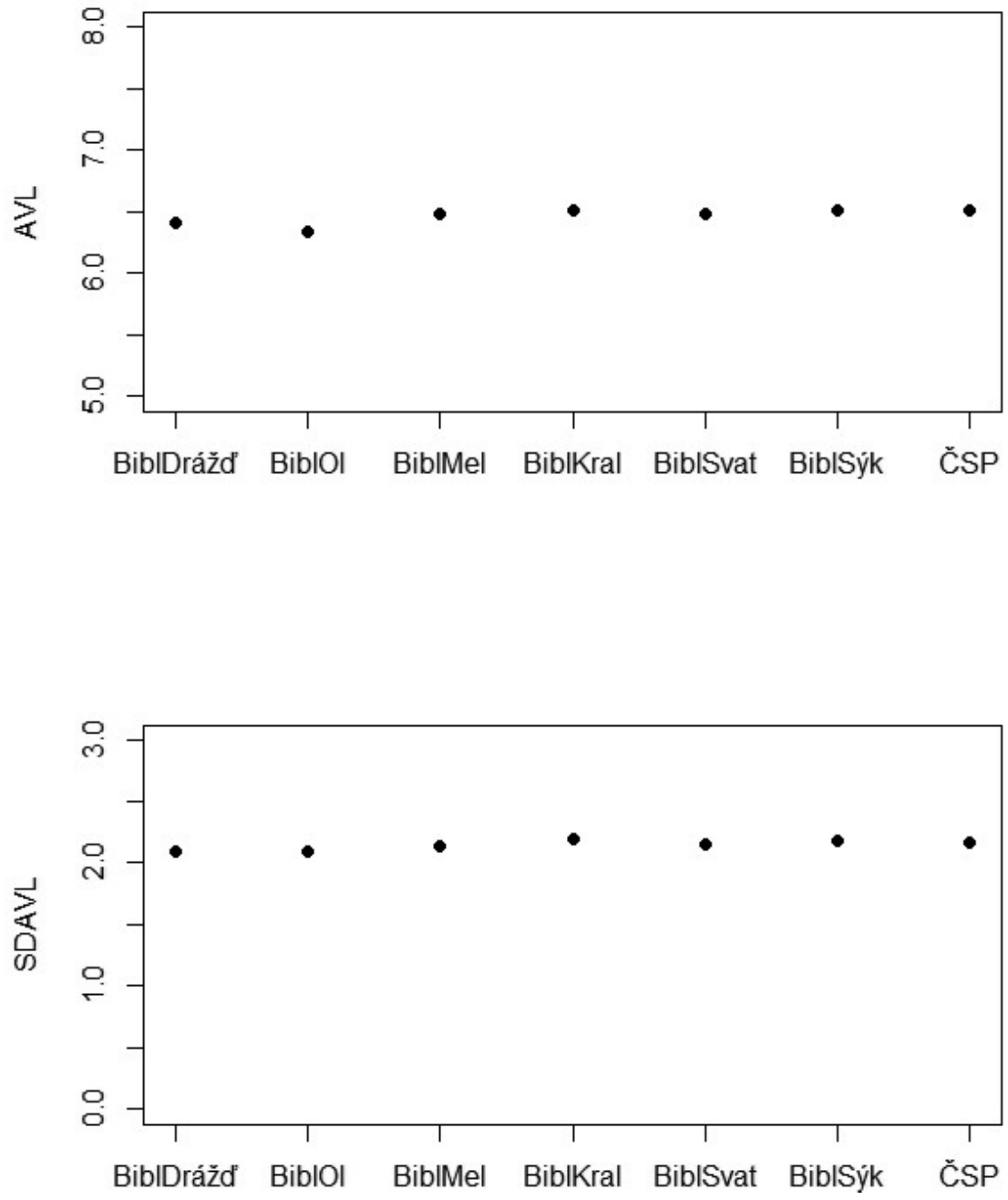

Fig. 2. The average type length (*AVL*) and the standard deviation of *AVL* (*SD*) in individual translations

There are no theoretical reasons to expect any direction in the historical development of the length (*N*) of particular translations, in our opinion. Of course, a variability of *N* can appear, as even a comparison of very short excerpts exemplifies, see (1) and (2) in Section 1. The data presented in Table 2 and Figure 3 reveal no single tendency in historical development. The biggest difference (between BiblOl and BiblSvat) is 735 tokens which means that the length of both texts differs by

approximately 4%. Further, there is an obvious difference between the two groups of texts. Specifically, a variability of *N* is evident among the first four translations, while in the group of the last three texts there are minimal differences of *N*. It can be explained as follows. In the 16th century, humanist scholars discussed the form of original Biblical texts which were used for translation. They were motivated by an effort to eliminate non-original parts which were added to the original during the Middle Ages. Since the end of the 16th century, there has been a relatively high consensus on the form of original Biblical text used for translation. Thus, the minimal variability of *N* in the last three translations in our sample can be interpreted as a consequence of this fact.

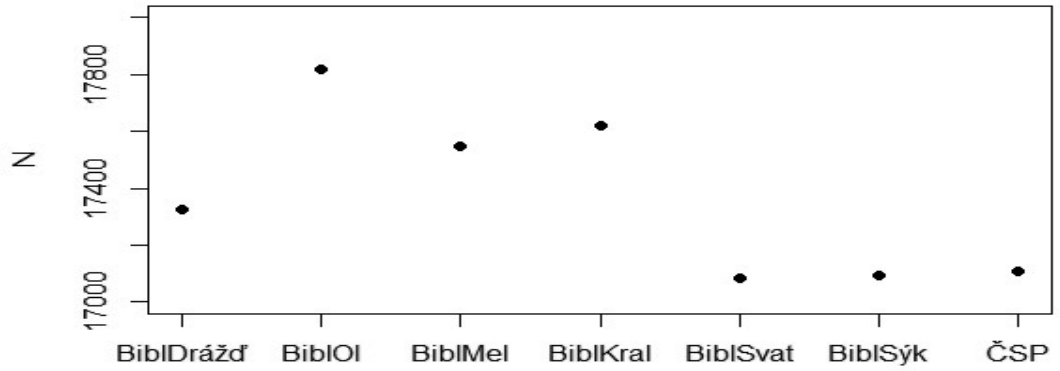

Fig. 3. The length (*N*) of individual texts

Bible	***PHL***	***MATTR***	***H***
BiblDrážď	0.154	0.781	9.981
BiblOl	0.133	0.766	9.826
BiblMel	0.139	0.775	9.896
BiblKral	0.141	0.781	9.936
BiblSvat	0.140	0.781	9.932
BiblSýk	0.144	0.784	9.955
ČSP	0.140	0.779	9.909

Tab. 3. The proportion of hapax legomena (*PHL*), the moving average type-token ratio (*MATTR*), and the entropy (*H*) of individual translations

The measurement of lexical diversity (sometimes called also vocabulary richness) captures how the writer (in our case the translator or translators) "manipulate" the vocabulary at his or her disposal to express the content. Obviously, there are several ways how to perform it. For instance, the usage of synonyms increases the lexical diversity, while a repetition of words (which one can use intentionally to ensure a high text cohesion) decreases it. There is no single method of measuring this property of the text. Here, we use three of them (*PHL*,

MATTR, *H*) to eliminate potential biases of particular methods. The results show (see Table 3 and Figure 4) that there are minimal differences among texts, regardless of the method.

At first sight, these results are rather surprising. Particular translations come from different periods, they were translated from different original texts (even from different languages), published by various churches for different recipients, the translators followed different translation strategies etc. Despite all these circumstances, the lexical diversity of all texts is almost identical. Most likely, there must be some dominant factor that eliminates the expected variability. In our opinion, the sacred nature of the gospel plays a decisive role here. Consequently, the translators strictly adhere to the lexical diversity of the original[3] and they do not dare to be creative in this regard.

4 CONCLUSION

In this study, we analyzed the historical development of several language and text properties of the Czech translations of the Gospel of Matthew to get a picture of their dynamics. We started with the most obvious one, the difference in word forms (*PIV*) in individual texts, and found out that the results are consistent with what is reported in the secondary sources. Further, the differences we detected in the text length (*N*) of particular translations can be explained if one realizes the historical context that influenced the translation process. The analysis of type length (*AVL*) brings the most surprising result we did not expect. This property seems to be extremely stable. Of course, this conclusion is based on the analysis of a very limited sample and we are aware that a much larger and diverse sample has to be examined to get a more reliable picture. Finally, the lexical diversity of individual texts is surprisingly stable in all texts. Here, we suggested the explanation which, however, must be confronted with other analyses of non-sacred texts.

Needless to say, the study is just a first step and have to be considered as a pilot study only. Only further research can corroborate (or falsify) the presented conclusions.

ACKNOWLEDGEMENTS

The study was supported by the project MUNI/FF-DEAN/1556/2019 “Vývoj pronominálních enklitik ‘mi’, ‘ho’, ‘mu’ ve starších českých biblích/The Development of the Czech Pronominal Enclitics ‘mi’, ‘ho’, ‘mu’ in Older Czech Bibles” (Pavel Kosek) and VEGA 2/0096/21 (Ján Mačutek).

[3] There is an open question if there are differences of lexical diversity in particular versions of original texts.

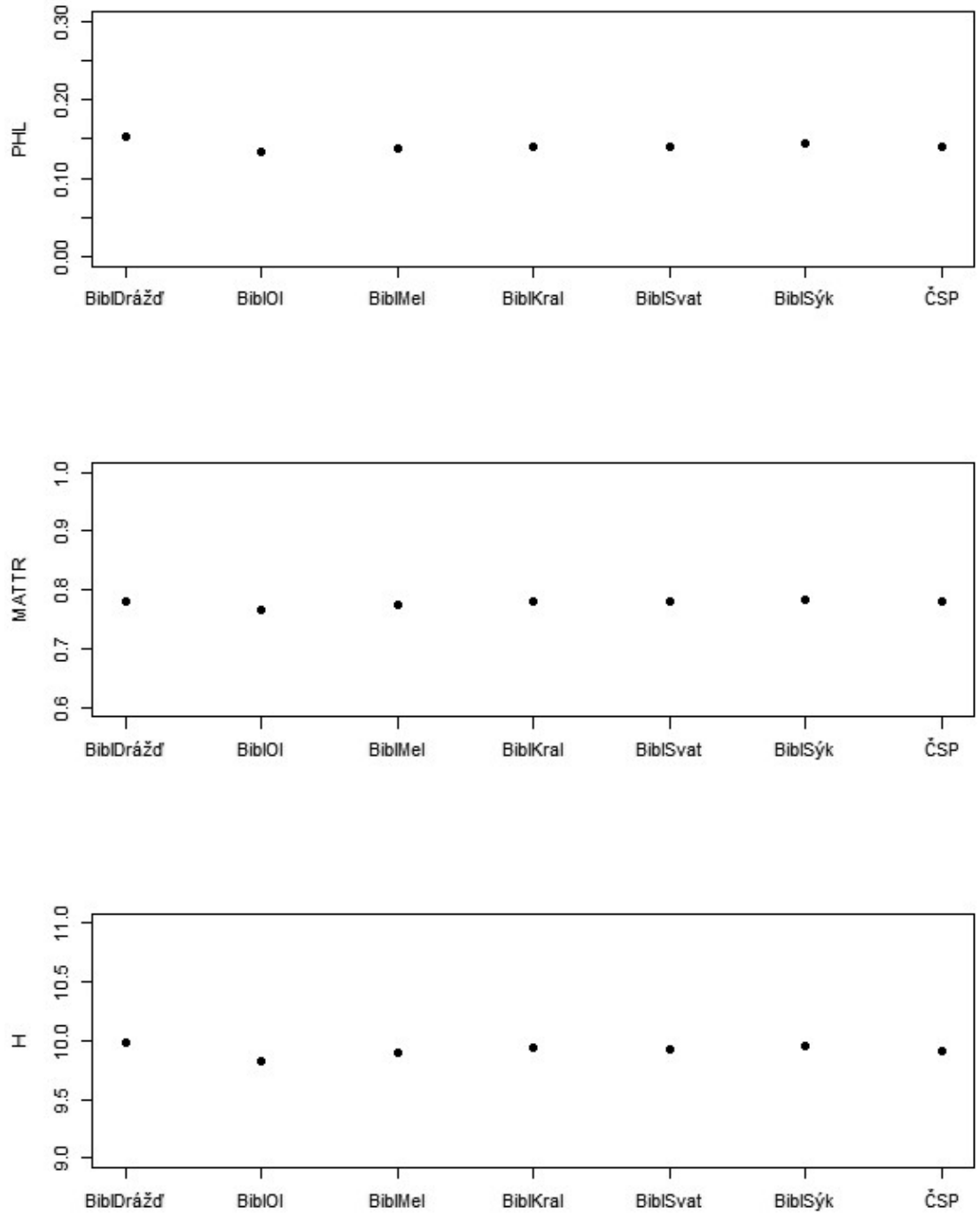

Fig. 4. The proportion of hapax legomena (*PHL*), the moving average type-token ratio (*MATTR*), and the entropy (N) of individual translations

References

[1] Čech, R., Kosek, P., Mačutek, J., and Navrátilová, O. (2020). Proč (někdy) nemíchat texty aneb Text jako výchozí jednotka lingvistické analýzy. Naše řeč, 103, pages 24–36.

[2] King James Bible Online. (2021). Available at: https://www.kingjamesbibleonline.org/.
[3] Kyas, V. (1997). Česká Bible v dějinách národního písemnictví. Praha: Vyšehrad.
[4] Vintr, J. (2008). Bible (staroslověnský překlad, české překlady). In L. Merhaut et al. (eds.), Lexikon české literatury. 4/II U–Ž, Dodatky A–Ř. Praha: Academia, pages 1882–1887.
[5] Cvrček, V., Čech, R., and Kubát, M. (2020). QuitaUp – a tool for quantitative stylometric analysis. Czech National Corpus and University of Ostrava. Available at: https://korpus.cz/quitaup/.
[6] R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/.