DOI: 10.24193/subbtref.69.II.02
Published Online: 2024-12-27
Published Print: 2024-12-27

M. PINTÉR Tibor[1] *– P. MÁRKUS Katalin*[2]*:*

Possibilities of Language Technology in Lexical Analyses of Canonical Hungarian Bible Translations[3]

Abstract.

The paper tries to attempt an unorthodox analysis of contemporary Hungarian Bible translations by using the software Sketch Engine. An important condition for computer-assisted linguistic analysis is a sufficient amount of textual data. In the case of Bible translations, "sufficient quantity" is easily attainable, since in this case a single translation represents the whole corpus. The corpus in this case is a properly annotated text: six Hungarian Bible translations. The paper highlights the fact that the computer-based text analysis can still reveal new features of Bible translations that are not found in the linguistic and hermeneutical analyses. In addition to the various layers of analysis of language use, it focuses on data like type-token characteristics, unigrams, bigrams, possible collocations, terms, orthographic error

[1] Associate Professor, Károli Gáspár University of the Reformed Church in Hungary; e-mail-address: m.pinter.tibor@kre.hu.

[2] Associate Professor, Károli Gáspár University of the Reformed Church in Hungary; e-mail-address: p.markus.kata@kre.hu.

[3] The research was funded by the project *Theolinguistic and Contrastive Linguistic Research / Teolingvisztikai és kontrasztív nyelvészeti kutatások* supported by Károli Gáspár University of the Reformed Church in Hungary.

types, and orthographic variations of the chosen translations. Analyses show new correlations between translations or prove the well-known connections between translations, denominations, and sacred language use. Data and statistics can be used for several purposes – here, to gain more knowledge about Hungarian Bible translations. Statistics do not influence the work of the translators, and no far-reaching conclusions can be drawn from them. But they do provide a basis for a more detailed interpretation of a kind that is not possible during human reading.

Keywords: *Bible translations, language technology, statistics, theolinguistics*

Introduction

This article attempts a relatively unorthodox analysis of some canonical Hungarian Bible translations. Using a well-established linguistic tool (Sketch Engine – https://www.sketchengine.eu), the texts of six Hungarian Bible translations (still in use today) are analysed. The analysis was primarily performed within the framework provided by the software, and at the same time (for example, in the study of name variations) the results of the software were processed by algorithms written by the authors. It is hoped that the computer analysis of the text will reveal features of the Bible translations not found in the linguistic and hermeneutical analyses. Apart from the many levels of language use analysis, each Bible translation's lexical properties and translation principles will be looked at, along with the features of sacral language use. The following analysis is based on the software Sketch Engine and the analysed results, which are used to display data and statistics about the Bible translations covered. It is important to note that the analysis is mostly statistics-based, not numerology – we are aware that the interpretation of the results has to be treated in its place. They do not influence the translation or the work of the translators, no far-reaching conclusions can be drawn from them; however, they do provide a basis for a more subtle interpretation that is not or only barely possible with the naked eye.

The Software

The linguistic use of language technology resources in Hungary is already diverse.[4] In addition to individual research projects, several research groups and universities are engaged in the channelling of info-communication and digital humanities into the human sciences and their teaching (in Hungary, the most renowned of these are Petőfi Museum of Literature or the Department of Digital Humanities at ELTE Eötvös Loránd University). The software used for analysis can be specifically designed for particular research projects and different focuses (e.g. the lexicographic software of the Termini Research Network by Tihamér Juhász, the BibAlign software by Zoltán Király Levente in the framework of the Unified Bible Reader project, or the Tibor M. Pintér's concordance software BibConcord), while more complex, more widely used software is probably more scientifically rewarding, as it can be used for several tasks (such as the use of ParaText software in "missionary" Bible translations). Reading, or Bible reading within the focus of our current study, is no longer confined to translations available on paper, the range of texts available online is also expanding,[5] facilitating faster and more multi-layered analyses or simply new ways of reading (e.g. parallel reading of the same verses).

[4] For example, only in relation to the research we have conducted: in the field of lexicography: M. PINTÉR, Tibor – P. MÁRKUS, Katalin – BENŐ, Attila (2023): Termini Online Hungarian Dictionary and Database (TOHDD): A Dictionary for Hungarian Varieties Spoken in the Carpathian Basin. In: *Acta Universitatis Sapientiae Philologica*. 15, 2. 166–181; P. MÁRKUS, Katalin – FAJT, Balázs – DRINGÓ-HORVÁTH, Ida (2023): Dictionary Skills in Teaching English and German as a Foreign Language in Hungary: A Questionnaire Study. In: *International Journal of Lexicography*. 36, 2. 173–194; concordance making: M. PINTÉR, Tibor (2022): Online bibliaolvasók szerepe és terminológiai megoldások keresése a bibliai konkordanciakészítésben. In: *Alkalmazott Nyelvtudomány*. 22, 1. 90–103; M. PINTÉR, Tibor – P. MÁRKUS, Katalin (2022): The Role of Online Bible Readers in Biblical Concordance Making. In: *Hungarian Studies Yearbook*. 4, 1. 183–196; online Bible readers: M. PINTÉR, Tibor (2021): Online segédletek a magyar nyelvű bibliafordítások olvasásához. In: *Modern Nyelvoktatás*. 27, 3–4. 43–57; corpora: M. PINTÉR, Tibor – P. MÁRKUS, Katalin (2021): Korpuszok a bibliafordításban mint a lexikológiai vizsgálatok eszközei. In: Fabiny, Tibor – M. Pintér, Tibor (eds.): *πῶς ἀναγινώσκεις; Hogyan olvasod? Felekezeteket összekötő Egyesített Bibliaolvasó (EBO) felé*. Budapest, Hermeneutikai Kutatóközpont. 72–89.

[5] For more, see, for example: M. PINTÉR 2021; KIRÁLY, Levente Zoltán (in press): Parallelizing Bible Texts. Developing the Database of the Unified Bible Reader online application. In: *Argumentum*.

In terms of text-processing capabilities, the Sketch Engine software, launched in 2004, was a new innovation, providing grammatical and lexical analysis and multilingual corpora, making it a very fast, powerful, and useful tool for those wishing to carry out linguistic analysis. One of the great (if not the greatest) advantages of the multifunctional software is that users can add their own texts to the quantitative paradigm, can upload their own texts and create corpora with grammatical analysis (and structural annotation). In other words, Sketch Engine is a comprehensive suite of text analysis tools designed to perform multi-functional text processing through an easy-to-use CQL query language. It also provides substantive quantitative data along the core functions while providing an opportunity for a possible extension of the qualitative analytical framework.

Several tools and operations are available in the fast and accurate analyses performed by the program. Below is a list of functions that can be used for text analysis in the software:[6]

• Concordance: searches words, phrases, tags, documents, text types, or corpus structures and displays the search results in context as a concordance, which can be sorted, filtered, and processed further to get the result needed. Complex searches, such as those with unspecific or optional criteria, are best performed using the CQL search option found on the advanced tab. It is a useful element in the construction of biblical concordances, since it not only sees the surface structure (character sequence) but also knows the morphological structures.[7]

• Word List: generates frequency lists of different kinds: 1. nouns, verbs, and other parts of speech; 2. words beginning, ending, containing certain characters; 3. word forms, tags, lemmas, and other attributes – or a combination of these three options. In addition, different frequency measures can be displayed in the word list, for example, absolute frequency (i.e. the number of occurrences of an item in the text) or frequency per million (i.e. the number of occurrences of an item per million tokens).

[6] For more on the program, see: JAKUBÍČEK, Miloš – KILGARRIFF, Adam – KOVÁŘ, Vojtěch – RYCHLÝ, Pavel – SUCHOMEL, Vít (2014): Finding Terms in Corpora for Many Languages with the Sketch Engine. In: *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics.* Gothenburg, Association for Computational Linguistics. 53–56; KILGARRIFF, Adam – BAISA, Vít – BUŠTA, Jan – JAKUBÍČEK, Miloš – KOVÁŘ, Vojtěch – MICHELFEIT, Jan – RYCHLÝ, Pavel – SUCHOMEL, Vít (2014): The Sketch Engine: Ten Years On. In: *Lexicography.* 1, 7–36.

[7] For possible difficulties in the construction of biblical concordances of Hungarian Bible translation, see M. PINTÉR 2022, M. PINTÉR – P. MÁRKUS 2021.

• Keywords: extracts terms for use in translation and interpreting; also extracts single-word and multi-word units that are typical of a corpus/document; as well as compares two corpora/documents/texts by identifying what is unique in the first corpus compared to the second. The result is divided into keywords (single-word items) and terms (multi-word items), displayed with links to the sentences in both the focus and the reference corpora. The tool can list the unique lexical items of a text, specific to that text (typical). A useful element is the search for word combinations, and multi-element key terms, where the frequency of a multi-element unit is given in relation to the elements of a reference corpus.

• Word Sketch: lists the word's collocates and other words in its surroundings. It can be used as a one-page summary of the word's grammatical and collocational behaviour. The results are organized into categories called grammatical relations (e.g. words that serve as an object of the verb, words that serve as a subject of the verb, and words that modify the word). Only the main word types (nouns, adjectives, verbs, and adverbs) are supported in most corpora. The collocations in the analysis are defined by rules specified in the sketch grammar.

• Word Sketch Difference: draws analogies via contrasting collocations. Three options are offered: 1. lemma: compares the use of two different lemmas through their collocates; 2. word forms: compares the use of two different word forms of the same lemma through their collocates; 3. subcorpora: compares the use of the same lemma in two different subcorpora of the same corpus through their collocates. Word collocates provide useful information on a word's usage, including subsenses, subject matter, connotations, and register. A thorough understanding of the variations in usage and meaning can be gained by comparing the collocations.

• Thesaurus: generates lists of words belonging to the same category (semantic field). The list is compiled based on the context in which the words appear in the selected corpus. Only the main word types (nouns, adjectives, verbs, and adverbs) are supported in most corpora. Because there is no manual work involved, synonym lists can be constructed for every word in the language as long as there are enough occurrences in the corpus. This is why synonym lists can be constructed for uncommon words that would not appear in typical thesauri.

- N-grams (multi-word expressions): builds frequency lists of sequences of tokens. The user has a variety of filtering options, including regular expressions, to determine which n-grams should have their frequency calculated. Words and lemma are the most commonly used attributes. Complex criteria for the n-grams that should be in the frequency list can be defined using regular expressions. The entire n-gram is interpreted as a single continuous string of characters, including spaces when utilizing regular expressions.
- Text type analysis: offers comprehensive statistics on the metadata of texts, documents, and elements imported into Sketch Engine.

Bible Translations

The primary criteria for the selection of the translations included in the analysis were that they should be biblical translations that are still in use and read widely today and that the translation should reflect as much as possible the current language use. Both requirements are in fact for comparability, since the translation remains current (not one that is no longer in use, whose linguistic elements are not in today's usage), and the similarity of language use and form makes the texts of the translations more comparable. István Kecskeméthy's translation differs somewhat from the above criteria, standing as a contrast to the translations showing the language usage of today. Several translations meet the above requirements (their parallel reading is offered, for example, by the Unified Bible Reader – Egyesített Bibliaolvasó, EBO, ebo.kre.hu – at Károli Gáspár University of the Reformed Church in Hungary), from which the more widely used canonical translations were chosen. Of these, the Catholic translations are represented by the *Bible of the Saint Stephen Association* (1973, hereafter SZIT), published by Saint Stephen Association (Szent István Társulat), and *Káldi-Neovulgate* (1997, hereafter KNV), published by Saint Jerome Catholic Bible Association (Szent Jeromos Katolikus Bibliatársulat). The Protestant translations are the following: Bible published by CE Koinonia Publishers (a translation by István Kecskeméthy, 2002 – hereafter KIF), the Newly Revised Bible of Károli published by Veritas Publishers (2020 – hereafter ÚRK), the Bible – The Hungarian Bible Society's New Translation (1990 – hereafter ÚF), published by John Calvin Publishing House of the Reformed Church in Hungary, and the Bible – The Hungarian Bible Society's Revised New Translation (2014 – RÚF), published by John Calvin Publishing House of the Reformed Church in Hungary and Hungarian Bible Society.

The following Bible translations are analysed in terms of their linguistic and language-use-related elements and do not deal with the theoretical or semantic analysis of the translation. The analysis is primarily concerned with the textual, character-based characteristics of the texts, which may serve as an interesting input for other analyses, but only provide information about the texts (i.e. their translation characteristics are not particularly important in this respect).

Language Use

The present analysis is based on five different Hungarian translations of the Holy Scripture. It aims to look for textual connections that would not be possible using paper-based translations. The study focuses more on the tool (the computer) than the text (translation). The question may be raised as to what extent this kind of analysis of specialized texts makes sense, how much it can help the reader or the translator, and even how much scientific value such analyses, which are desacralized, have. The scientific nature is justified by the analytical framework and the fact that it is part of the scientific discourse analysing biblical translations. The research of the analysed Bible translations in this framework can be understood as a text-centred rather than an inspiration-oriented or hermeneutical analysis. This kind of approach does not imply a complete desacralization of the texts (translations) – it raises the question of how open the text(s) of the Bible are to a possible secularized analysis and interpretation.

Bible translations are *ab ovo* sacred texts. However, a secular analytical framework anticipates the question: what makes the Bible and other inspired texts sacred? The sacrality of biblical translations, and the extent to which they are sacred, can be approached from several different perspectives. From a hermeneutical point of view, it can be defined by the following criteria. According to one approach, a sacred text is one that: 1. considers itself to be divinely inspired; 2. is the bearer of divine revelation; 3. carries a coded, hidden, "secret" content, a message whose interpretation is not clear; 4. its interpretation requires a privileged interpreter; 5. has life-forming, life-determining content; 6. serves as the basis for a religious rite; 7. evokes the divine presence.[8] In addition to the challenges mentioned above, the translation of sacred texts also raises

[8] Cf. DETWEILER, Robert (1985): What Is a Sacred Text? In: *Semeia*. 31, 213–230.

specific problems and confirms its specialized nature. The extent to which the awareness and control, regulation of the translators change the level of inspiration of the original texts (cf. the requirements for regulated and unregulated translations of sacred texts) must not be neglected.[9] In the case of regulated translation, for the inspiration of the translation, it is important to know who is translating, what and what texts are considered as source texts, what translation principles and methods are followed, what the target audience is, and who has reviewed and commented the completed texts before finalizing the translation. It is a fundamental principle of the translation of sacred texts that the inspiration of the translation can only be preserved if a conscious (and initiated) group of translators translates in such a way that the original texts (not necessarily the source texts from the point of view of translation) remain hidden from the "profane" (external) readers or the original texts remain in a kind of "coded" writing, or the texts are read and interpreted within the framework of a specific group. And perhaps that could be provided only by Bible scholars (that is why Christiane Nord claims Bible translation as a work "in the hands of theologians").[10]

In determining the sacrality of texts, not only the content but also the language is a determining factor (in this context, it is worth mentioning that the "very characteristic language" of the Károli translation is always emphasized when Protestant translations are examined). The sacrality of language and language use is in this context a sophisticated correlation between several linguistic and hermeneutic features. As Gergely Hanula argues, sacred language use is characterized by the following properties of language and language use, which are very similar to Detweiler's hermeneutical approach: 1. permanence (in the case of sacred texts and their translation, the subject and the object of the texts are always the same), 2. limitation (lexical, syntactic, and semantic constraints in the languages concerned), 3. vagueness of meaning (either in terms of wording or in terms of content, the inconceivability of the representation), 4. indirectness (the message is not conveyed by the sender but by an intermediary person or persons), and 5. the renunciation of individual intention (linguistic legitimation of the cessation of the self – the person who speaks during the ritual is mediating, not

[9] Cf. NAUDÉ, Jacobus (2010): Religious Translation. In: Gambier, Yves – Van Doorslaer, Luc (eds.): *Handbook of Translation Studies 1.* Amsterdam, John Benjamins. 285–293.

[10] See NORD, Christiane (2016): Function + Loyalty: Theology Meets Skopos. In: *Open Theology*. 2, 1. 566–580.

expressing themselves). These main linguistic features together create the sense of strangeness and inspiration essential in sacral language.[11]

In addition to archaicism, the language of biblical translations is characterized by biblical vocabulary and, in a wider perspective, by specific biblical language use (e.g. specific meanings of lexemes and collocations, the use of unique syntactic units different from colloquial usage, stylistic elements), which is a feature of sacral language use in general. Thus, the autonomy of sacral language as a linguistic variety characterizing the use of special texts also reinforces the specialized linguistic character of Hungarian Bible translations (this is confirmed by international and also Hungarian theolinguistic research). The specialized nature of the text is justified on the one hand by the vocabulary, the style (and even the diversity and richness of styles), and the thematic-cultural context and on the other hand by the theoretical questions and problems related to the creation and translation of texts (the sacral nature of sacred texts determines the use of language, which thus represents a specific register; for the Hungarian context see studies published in the 1. thematic volume in 2024 of the linguistic journal *Alkalmazott Nyelvtudomány* [Hungarian Journal of Applied Linguistics]).

From this brief theoretical outline, it is apparent that the specificity of biblical language can be characterized in several ways: for example, as it is described by Jan de Waard and Eugene Nida as a language that deals with supernatural phenomena that have no established linguistic toolkit and that reflects transcendent experiences that conventional language seems incapable of describing.[12]

In relation to sacred texts as well as sacral and biblical language, it is worth mentioning the language of simplified or easy-to-read translations, which raises the question of the extent to which the secular interpretation and use means their desacralization. There is certainly a need for popular or common language translations of the Bible.[13] Thus, perhaps inspiration is not reached only through the use of sacral language or special linguistic features.

[11] HANULA, Gergely (2016): *Anyaszentnyelvünk. A „szent nyelvek” és a fordítás.* Budapest, Argumentum Kiadó – ELTE BTK Vallástudományi Központ Liturgiatörténeti Kutatócsoport – Pápai Református Teológiai Akadémia. 94–95, 102.

[12] DE WAARD, Jan – NIDA, Eugene A. (1986): *From One Language to Another: Functional Equivalence in Bible Translating*. Nashville, Nelson.

[13] Cf. WONDERLY, William L. (1970): Some Principles of "Common Language" Translation. In: *The Bible Translator*. 21, 3. 126–137.

The Framework for Analysis

An important prerequisite for computer-assisted linguistic analysis is a sufficient amount of linguistic data. In the case of Bible translations, "sufficient quantity" is easily attainable, since in this case a single translation is in fact the whole underlying corpus. The corpus in this case is a properly annotated text, and for the aims of this analysis, we rely only on the bibliographic and grammatical analysis. The advantage of computer-based analysis is that, through targeted research, it is possible to obtain a sufficient quantity and quality of data in a relatively short time – which may be used for more superficial but also more in-depth analyses.

Among the analysis options offered by the program, the following features of Sketch Engine were used for the study: *word list, 2-gram, keywords*. The analyses can provide insights into the lexical properties of each Bible translation: the most and least frequently used word forms (word frequency); the most common collocations and structures; typical expressions and phenomena specific to the translation. The comparatively presented analyses can be performed in a more detailed and refined context on each Bible translation separately, providing a framework for further in-depth computer analyses of each Bible translation.

The Bible translations used for the analysis were source texts from the Unified Bible Reader (ebo.kre.hu) project.

Analyses

Word Count

The easiest of the software analyses is the comparison by word count (or number of tokens). The program displays the number of character strings between textual delimiters, which is in this case the space (before the analysis, non-alphabetic characters for syntax, punctuation marks are removed). The word count comparison can be misleading, as the basis of comparison are orthographic words (words between spaces). Analytical and synthetic rendering of words (compound word vs syntactic structure) can lead to significant differences in the results. In case of the present research, the same or at least not significantly different source text(s) are used, therefore the similarities and differences in word count cannot be attributed to potentially different editing methods (differences can be within Protestant and Catholic translations).

The most noticeable difference between Protestant and Catholic Bible translation is the word count (further other differences are going to be presented). This is not, however, a consequence of differences in translation, translation methods, or source texts but rather evidence of differences in the content of the translations. It should be noted that the so-called deuterocanonical books by the Catholic canon are not part of the Protestant translations, which means that they are only found in the Catholic translations (here: SZIT and KNV). Consequently, the word counts of the two Catholic translations (SZIT and KNV) are significantly higher than those of the Protestant translations. The analysed translations were manually cleaned, removing the number of verses, punctuation, and non-letter characters from the texts (some of them were text conversion errors, hyphens, and other signs). After comparing the clean texts, the following results were obtained:

Table 1. *Word count of the Bible translations analysed*

	RÚF	ÚF	ÚRK	KIF	SZIT	KNV
token	549 273	536 981	537 497	556 614	612 740	630 580
type	56 907	57 183	57 919	59 437	66 897	66 320

These figures suggest (as was already evident) that there is no significant difference between the word counts of ÚF and RÚF, and that the ÚRK word count is closest to ÚF. It is interesting to note that the oldest translation examined does not differ significantly in word count from modern translations, although its absolute word count is slightly higher.

Errors

When it comes to Bible translations, writing about errors is not the most correct thing to do. In our analytical framework, the search for errors is more about analysing the performance of the analyser than about qualifying the text of the translations (we do not aim at finding errors, nor do we feel entitled to mention "error" in the context of

Bible translations; however, in the context of analysing texts with a particular style and language, the analyser can be "analysed" in this respect). After the analysis, words labelled by the analyser as "UNKNOWN" can be divided into two categories: a) words that the analyser does not recognize are errors and b) incorrect words that are incorrect only for the analyser (incorrect recognition).

In many ways, it is reassuring that the six translations analysed contain a negligible number of errors: reassuring because it indicates that the analyser works properly. Although a certain degree of archaism is typical of even the most modern translations, it also indicates that the relatively long texts contain few spelling mistakes and mispronunciations. In this respect, it is worth noting that the analyser uses a character-based text recognition algorithm, i.e. it does not distinguish between homonymous forms and cannot handle the categories of Hungarian orthography: writing a word as a single one or presented in a syntactic structure, thus it only filters out words that cannot be analysed.

The texts analysed show the following number of errors:

Table 2. *Number of errors*

RÚF	ÚF	ÚRK	KIF	SZIT	KNV
7	11	23	51	19	16

According to the examples below, the words considered to be incorrect can be divided into the following major groups:

- typical biblical words that do not have a homonymous form with which the analyser is familiar (e.g. *hínnyit*, *Eltekét*);
- archaic words (e.g. *szőlőtő*) which the analyser is not familiar with, or which could be considered as a misspelling (e.g. *csalárd* as a misspelling of *család*);
- writing numerals (e.g. *kétezerhétszáz*, *ötezernégyszázat*);
- typographical error (*harmickét*, *nEgyEdik*);
- misinterpretation of the superscript as a number (*anyád52*, *azokat70*) – these are not translation errors but rather structural, technical errors.

The following examples also show that two distinct types of error characterize two Bible translations: ÚRK is characterized by typographical errors; on the other hand, KIF is characterized by a misinterpretation of the superscript.

Table 3. *Words found to be incorrect by Sketch Engine*

RÚF	Eltekét, Gallió, Le, bat, et, kétezerhétszáz, szőlőtőt
ÚF	Eltekét, Gallió, Le, bat, et, ezerízig, hétezerhétszáz, kétezerhétszáz, kétezerszázhetvenkét, szőlőtőt, ötvenháromezernégyszázat
ÚRK	Eltekét, Gallió, HarMadik, Le, Rabsakét, al, bat, ben, csalárdot, da, et, ezerízig, ezret-ezret, gyé, harmadízig, harmichárom, harmickétezer-ötszáz, la, los, minc, nEgyEdik, szőlőtőt, té
KIF	Eltekét, En, Gallió, Le, On, Rabsákét, anyád52, azokat70, beborította18, betöréskor10, bériek87, cserbenhagyjam37, dágon16, emberek9, ezerízig, fajzat18, fel32, felemeli30, fiai9, fogyjon6, hozzá33, húszezerkétszáz, ipának2, irányban18, kerek63, kilenvenkilenc, kiontott9, kiosztotta94, királyának16, legyen53, megvetetted48, megvénhed, melyek40, negyven67, neveztetett48, nyakán54, orgyilkos156, paráznanőt, ruhát63, régebb, szállnak49, százennyit, szőlőtőt, tartott9, vállrakötőt, ébenfát58, úgy75, út14, úton6, útra1, őelőtte76
SZIT	000-et, 16d, Eltekét, Gallió, Le, On, bat, es, háromezerhatszáz, háromezerhatszázat, hétezerhétszáz, kétezerhatszáz, kétezerhatvanhét, kétezerhétszáz, kétezerötvenhat, száznyolvanezer, ítéltd, összefűzköd, ötezernégyszázat
KNV	En, Gallió, Le, Rábsakét, an, ezerízig, ezret-ezret, gabonaharmadot, háromezerhuszonhárom hínnyit, kijöttöd, kétezernégyszáz, négyeze-rötszáz, négyezerhatszáz, négyezerötszáz, szőlőtőt

It should also be mentioned that the analyser is also capable of hallucinating or over-analysis, where correct forms – usually in nominative case – are analysed as derived forms: for example, the noun *Lélek* 'soul' beginning with a capital letter as the plural of *Lél*; the verb form *Intelek* 'I admonish you' as the plural of the noun *Intel*; the form *Kittim* 'isle of Cyprus or its inhabitants' as the possessive case of the verb *Kitti* or the form *Ladán* 'a descendant of Ephraim, and an ancestor of Joshua' as the locative case of *Lada*.

2-gram

A fundamental element of language processing analyses, such as language modelling or even the analysis of collocation candidates, is the frequency list of two-element items (the list may, however, contain units of several elements, depending on the morphological and syntactic properties of the language). A 2-gram or bigram looks at the sequence of words in a text that are next to each other, always including a sequence of 2–2 words, meaning that an orthographic word is included by two bigrams (with the word before it as the 2nd word and the word behind it as the 1st word). This kind of distribution shows which words are most frequently next to each other in a text, as well as whether a word combination element is more likely to be a first or a second item – showing, for example, the relationships between words).

Within the given analytical framework, we will examine which two-element word combinations occur most frequently in each Bible translation. Due to the morphological properties of the Hungarian language, we assume that they will be primarily personal nouns and adjectival structures.

In this study, word roots are examined, not word forms, since we do not intend to investigate the morphology of Hungarian as a target language but the frequency of lexemes in the source language. Since we are unfamiliar with the translation process, the scopus, or the translation brief, we will not delve into the specifics of a deeper investigation of the discrepancies in each translation. Even before the computer analysis of the "most common words" in Bible translations, it is assumed that the most frequently used words will be those specific to the Christian religion – so, the analysis can only reveal something new and scientifically significant about the possible differences between the translations. Accordingly, rather than evaluating the bigrams of separate Bible translations, it is better to consider them in relation to one another.

The 2-grams in *Table 4* contain the types that occur at least 100 times in each translation (otherwise the table would be too long). The table shows the pre-analysis hypothesis: i.e. no significant difference is expected between translations. However, the bolded words in the table show the characteristics of a particular translation. Types in bold characterize only one or a few translations (assuming that the first 100 occurrences characterize), hence focus will be placed only on those types.

Table 4. *Bi-grams for each translation*

RÚF		ÚF		ÚRK		KIF		SZIT		KNV	
izráel fiú	616	izráel fiú	605	izráel fiú	594	izráel fiú	632	izrael fiú	717	izrael fiú	611
úr ház	227	úr ház	222	úr ház	233	égő áldozat	266	seregek úr	256	**úr isten**	283
izráel király	214	izráel király	210	seregek ura	210	úr ház	238	izrael isten	207	seregek úr	250
úr ige	208	seregek ura	206	izráel isten	198	**úr szó**	231	izrael király	204	**elégő áldozat**	228
izráel isten	204	izráel isten	193	**egyiptom föld**	190	seregek úr	230	júda király	186	izrael isten	214
seregek ura	195	júda király	168	izráel király	175	**úr mózes**	209	jézus krisztus	161	úr ház	205
júda király	180	kijelentés sátor	141	jézus krisztus	151	**izráel isten**	203	úr templom	153	izrael király	196
kijelentés sátor	145	jézus krisztus	139	júda király	149	jézus krisztus	195	**egyiptom föld**	139	júda király	190
harci kocsi	144	harci kocsi	136	**való áldozat**	144	izráel király	164	izrael ház	132	**egyiptom föld**	179
jézus krisztus	144	úr szín	122	Gyülekezet sátor	135	gyülekezet sátor	136	úr szó	124	**úr mózes**	161
úr szín	127	izráel ház	120	izráel ház	122	júda király	136	**megnyilatkozás sátor**	112	jézus krisztus	157
háza nép	124	háza nép	110	**elégő áldozat**	112	izráel ház	130	babilon király	111	izrael ház	132
izráel ház	121	úr ige	105	**úr szó**	109	**egyiptom ország**	125	**úr mózes**	104	babilon király	130
		babilónia király	103	**bűnért való áldozat**	107	**királyi szék**	102			való áldozat	117
				háza nép	107						
				szent hely	104						

This suggests that ÚRK, KIF, SZIT, and KNV contain specific word combinations unique to those translations (e.g. *való áldozat*, *szent hely* – ÚRK, *egyiptom ország* – KIF, *megnyilatkozás sátor* – SZIT, KNV – *úr isten*). In addition, it is also interesting to look at the frequency of the individual items. From the latter, the "main character" or "main characters" of the translations become clear even without any background knowledge. Some lexical differences between the translations are thus revealed (e.g. *kijelentés* –

megnyilatkozás, *ige – szó*), and the interrelationship between the translations is also more clearly shown. In this respect, the relationship between ÚF and RÚF and SZIT and KNV seems clear; however, the ÚRK, published as a translation of Károli, is more distant from ÚF and RÚF and also from KIF. Although the earliest translation, KIF, appears to be different from the other translations (*égő – elégő*, *szó – ige*), it becomes apparent that this difference is negligible compared to the other translations when one is aware of the relatively large temporal difference.

Term

Another indicator of lexical frequency is the list of typical words used in the text. For Sketch Engine, a word becomes a term (i.e. a typical word in a text set) if it is under-represented in another, general corpus (of which there are plenty in the software) – i.e. a word in a given text set is considered typical if it is under-represented in other general corpora managed by the software.

As in the case of the bigrams, typical words in the Bible are likely to be related to its content without prior analysis, i.e. common words that are not characteristic of a single book but define the Old and New Testaments separately or simultaneously occur frequently in them (as opposed to other non-biblical texts). Thus, as in the case of the bigrams, the usefulness of the analysis is not in the relationship between translations but in their comparison.

Table 5 lists the words in the translations that occur at least a thousand times in each Bible translation. The direct relationship between the ÚF and the RÚF is clear also in this respect, as is the relationship of the ÚRK to the other translations. The ÚRK is also closer to the other three translations than to the new Protestant translations (ÚF, RÚF) examined in terms of the characteristic words.

This is not surprising, however, since in the preface to the translation the translators indicated that the translation's floridity and the text's flavour might make the text more archaic, and they also move away from the explicit realization of linguistic modernity, i.e. they indicate explicitly that the ÚRK is more like the older translations. The translations analysed do not show significant differences in terms of terminology (which is to be expected). Still, it can be noted that the most typical words in the translations analysed are *Úr* 'Lord', *Isten* 'God', *fiú* 'son', and *király* 'king'.

Table 5. *Terms specific to each translation*

RÚF		ÚF		ÚRK		KIF		SZIT		KNV	
úr	6432	úr	6082	úr	6135	úr	8234	úr	8374	úr	8473
isten	4327	isten	4230	fiú	4445	fiú	4829	fiú	4993	isten	5081
fiú	4272	fiú	4136	isten	4413	isten	4306	isten	4935	fiú	4926
király	2896	király	2800	király	2731	király	2972	király	3173	király	3143
nép	2659	nép	2593	nép	2571	ti	2900	nép	2719	föld	2779
ti	2650	ti	2564	ti	2570	izráel	2554	izrael	2689	ti	2721
izráel	2491	izráel	2449	**föld**	2562	nép	2114	**föld**	2510	izrael	2610
azután	1198	dávid	1118	izráel	2445	atya	1500	ti	2423	nép	2399
dávid	1136	azután	1060	**íme**	1133	**mondván**	1250	atya	1169	atya	1140
				szolga	1105	dávid	1135	dávid	1115	dávid	1059
				atya	1043	cselekszik	1008	**szolga**	1001	**íme**	1032
				dávid	1042						
				cselekszik	1036						

Spelling Variants

The problem of the spelling of biblical names is not a new one and is by no means unique to Hungarian translations: the variability of names appears already in the Septuagint, and their transcription is not stable in the languages used today.[14] As Zoltán Kustár explains, the transcription of geographical names and personal names is a perennial challenge for Bible translators.[15] Yet the link between the source and target language texts is not only a conscious intention of the Bible translator: in addition to the hidden or more public influence of linguistic ideologies, translators must pay attention to the expectations of the readers, the general spelling rules and spelling conventions of the target language, and the accepted or established (even regulated) writing conventions of the churches.

[14] For more information, see: KUSTÁR, Zoltán (2015): A bibliai héber nevek megjelenítése a nemzeti bibliafordításokban, különös tekintettel a legújabb protestáns bibliafordításainkra. In: *Névtani Értesítő*. 37, 25–32; KRAŠOVEC, Jože (2010): *The Transformation of Biblical Proper Names*. New York – London, T & T Clark.

[15] KUSTÁR 2015, 30.

A Bible translation is a huge undertaking, usually carried out by translators and biblical scholars (sometimes with the help of linguists). Compared with other translations and primary texts, the target language text is revised continuously so that in modern translations linguistic and spelling errors are minimal (as shown above). The spelling differences are differences in the written form of a proper name, which can be linked to the different denominations and their Bible translations. These variations are already found in the Septuagint and the Vulgate, demonstrating the power of transliteration and its role in languages. Thus, the actual spelling variants are known in advance, but within a translation, spelling variants of the same proper name can also appear. In the following, we will only deal with (or give a taste of) those cases that fall into the latter category.

The following are some examples based on a morphosyntactic analysis of Sketch Engine. The examples are based on a filtered list of nouns with capital initials generated by the program, namely proper nouns that have appeared in several versions of a single translation and refer to the same denotatum (Protestant–Catholic variants are not discussed here). Since a Bible translation is the result of a long process and several rounds of checking, it can be assumed that there are not many inconsistencies in the name elements of each translation. The fact that the examples below are complementary is confirmed by the number of occurrences: one or two occurrences are presumably "left" in the translations (similar "left in" occurrences are known, for example, in the German–Hungarian and English–Hungarian dictionaries of Akadémiai Publishers, but also in Osiris Orthography dictionary, which may have a function – texts stored on computers are easily copied, so if the "incorrect" text were to occur in an unsolicited place, there could be copyright consequences).

From the examples below, it appears that the variability of the name variants is mainly in the translations of SZIT and KNV. It is important to note that since we have not reviewed the entire list of almost 20,000 lines, the examples below are only a sample (the number of occurrences is given in brackets):

RÚF:
Abdeél (2) – Abdíél (1)

KIF
Adbéél (1) – Adbeél (1)
Izráel (2554) – Izrael (5)

SZIT:
Baal (84) – Baál (2)
Benjamin (176) – Benjámin (1)
Micha (30) – Mika (6)
Adullam (4) – Adullám (2)

KNV:
Ráchel (35) – Ráhel (3)
Rebekka (24) – Rebeka (1)
Zabdiel (1) – Zabdiél (1)

The Role of Language Technology in the Analysis of Bible Translations

The benefit of the above analysis is that it makes visible properties of texts that would be difficult to detect with the human eye while reading. The rapid comparative analysis of the lexical properties of the text focused primarily on quantifiable elements. The typical word usage of each translation makes the translations similar or even different. In this respect, the list of typical words for each translation, as well as the analysis of word structures, collocations, and candidate collocations for each translation, showed interesting results. The number of words and the number of word fragments in the translations do not influence the reading of the Scriptures, but they do characterize the translators' use of language, their use of translation procedures and their reliance on translation solutions. The variations in proper names of each translation could be revealed by lengthy manual analysis, while targeted searches could yield more accurate and faster results.

The biggest advantage of computer-generated texts is that character-based analyses can be performed faster and more accurately than on printed texts. In addition to various grammatical analyses, new layers and contexts of texts can be revealed to the researcher or reader. Sketch Engine, as the most widely used linguistic analyser, is just one way of carrying out general analyses. Besides the analytical framework mentioned in the introductory sections, morphosyntactic analysis also allows for other, derived analyses – in this case, however, the output of the software will be the input text of another (even locally created) analyser. Digital literacy is now an indispensable tool for Bible readers, concordance makers, and automatically generated dictionaries. Such analyses can be useful not only to understand the text but also to grasp the properties and peculiarities of the translation. In this way, the preparation of Bible readers and linguistic or hermeneutical research can go hand in hand, facilitating the work of each field.

References

Alkalmazott Nyelvtudomány [Hungarian Journal of Applied Linguistics] (2024): *Teolingvisztikai különszám*.

DE WAARD, Jan – NIDA, Eugene A. (1986): *From One Language to Another: Functional Equivalence in Bible Translating*. Nashville, Nelson.

DETWEILER, Robert (1985): What Is a Sacred Text? In: *Semeia*. 31, 213–230.

HANULA, Gergely (2016): Anyaszentnyelvünk. A „szent nyelvek" és a fordítás. Budapest, Argumentum Kiadó – ELTE BTK Vallástudományi Központ Liturgiatörténeti Kutatócsoport – Pápai Református Teológiai Akadémia.

JAKUBÍČEK, Miloš – KILGARRIFF, Adam – KOVÁŘ, Vojtěch – RYCHLÝ, Pavel – SUCHOMEL, Vít (2014): Finding Terms in Corpora for Many Languages with the Sketch Engine. In: *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Association for Computational Linguistics. 53–56. https://doi.org/10.3115/v1/E14-2014.

KILGARRIFF, Adam – BAISA, Vít – BUŠTA, Jan – JAKUBÍČEK, Miloš – KOVÁŘ, Vojtěch – MICHELFEIT, Jan – RYCHLÝ, Pavel – SUCHOMEL, Vít (2014): The Sketch Engine: Ten Years On. In: *Lexicography*. 1, 7–36. https://doi.org/10.1007/s40607-014-0009-9.

KIRÁLY, Levente Zoltán (in press): Parallelizing Bible Texts. Developing the Database of the Unified Bible Reader online application. In: *Argumentum*.

KRAŠOVEC, Jože (2010): *The Transformation of Biblical Proper Names*. New York – London, T & T Clark.

KUSTÁR, Zoltán (2015): A bibliai héber nevek megjelenítése a nemzeti bibliafordításokban, különös tekintettel a legújabb protestáns bibliafordításainkra. In: *Névtani Értesítő*. 37, 25–32. https://doi.org/10.29178/NevtErt.2015.2.

M. PINTÉR, Tibor (2021): Online segédletek a magyar nyelvű bibliafordítások olvasásához. In: *Modern Nyelvoktatás*. 27, 3–4. 43–57. https://doi.org/10.51139/monye.2021.3-4.43.57.

(2022): Online bibliaolvasók szerepe és terminológiai megoldások keresése a bibliai konkordanciakészítésben. In: *Alkalmazott Nyelvtudomány*. 22, 1. 90–103.

M. PINTÉR, Tibor – P. MÁRKUS, Katalin (2021): Korpuszok a bibliafordításban mint a lexikológiai vizsgálatok eszközei. In: Fabiny, Tibor – M. Pintér, Tibor (eds.): *πῶς ἀναγινώσκεις; Hogyan olvasod? Felekezeteket összekötő Egyesített Bibliaolvasó (EBO) felé*. Budapest: Hermeneutikai Kutatóközpont. 72–89.

(2022): The Role of Online Bible Readers in Biblical Concordance Making. In: *Hungarian Studies Yearbook*. 4, 1. 183–196. https://doi.org/10.2478/hsy-2022-0009.

M. PINTÉR, Tibor – P. MÁRKUS, Katalin – BENŐ, Attila (2023): Termini Online Hungarian Dictionary and Database (TOHDD): A Dictionary for Hungarian Varieties Spoken in the Carpathian Basin. In: *Acta Universitatis Sapientiae Philologica*. 15, 2. 166–181. https://doi.org/10.2478/ausp-2023-0023.

NAUDÉ, Jacobus (2010): Religious Translation, In: Gambier, Yves – Van Doorslaer, Luc (eds.): *Handbook of Translation Studies 1*. Amsterdam, John Benjamins. 285–293. https://doi.org/10.1075/hts.1.rel3.

NORD, Christiane (2016): Function + Loyalty: Theology Meets Skopos. In: *Open Theology*. 2, 1. 566–580. https://doi.org/10.1515/opth-2016-0045.

P. MÁRKUS, Katalin – FAJT, Balázs – DRINGÓ-HORVÁTH, Ida (2023): Dictionary Skills in Teaching English and German as a Foreign Language in Hungary: A Questionnaire Study. In: *International Journal of Lexicography*. 36, 2. 173–194. https://doi.org/10.1093/ijl/ecad004.

WONDERLY, William L. (1970): Some Principles of "Common Language" Translation. In: *The Bible Translator*. 21, 3. 126–137. https://doi.org/10.1177/000608447002100303.

Bible Translations

KIF = Kecskeméthy (Csapó) István (1931/2002): *Biblia* [Bible]. Kolozsvár, CE Koinónia Kiadó.

KNV = (based on the rev. by György Káldi) (1997): Ó- *és Újszövetségi Szentírás a Neovulgáta alapján* [Káldi-Neovulgate]. Budapest, Szent Jeromos Katolikus Bibliatársulat.

RÚF = 2014: *Biblia – Revideált új fordítású* [Revised New Translation]. Budapest, Kálvin Kiadó – Magyar Bibliatársulat.

SZIT = Rózsa Huba (EIC) (1973): *Biblia – Ószövetségi és Újszövetségi Szentírás* [Bible of the Saint Stephen Association]. Budapest, Szent István Társulat.

ÚF = 1990: *Biblia – Új protestáns fordítás*. 1. revízió [New Translation]. Budapest, Kálvin Kiadó – Magyar Bibliatársulat.

ÚRK = 2020: *Újonnan Revideált Károli-Biblia* [Newly Revised Bible of Károli]. Budapest, Veritas Kiadó.