Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4270–4274
Marseille, 20-25 June 2022

From Pattern to Interpretation. Using Colibri Core to Detect Translation Patterns in the Peshitta.

Mathias Coeckelbergs
Université libre de Bruxelles
Franklin Rooseveltlaan 50, 1050 Brussels
Mathias.Coeckelbergs@ulb.be

Abstract

This article presents the first results of the CLARIAH-funded project 'Patterns in Translation: Using Colibri Core for the Syriac Bible' (PaTraCoSy). This project seeks to use Colibri Core to detect translation patterns in the Peshitta, the Syriac translation of the Hebrew Bible. We first describe how we constructed word and phrase alignment between these two texts. This step is necessary to successfully implement the functionalities of Colibri Core. After this, we further describe our first investigations with the software. We describe how we use the built-in pattern modeller to detect n-gram and skipgram patterns in both Hebrew and Syriac texts. Colibri Core does not allow the creation of a bilingual model, which is why we compare the separate models. After a presentation of a few general insights on the overall translation behaviour of the Peshitta, we delve deeper into the concrete patterns we can detect by the n-gram/skipgram analysis. We provide multiple examples from the book of Genesis, a book which has been treated broadly in scholarly research into the Syriac translation, but which also appears to have interesting features based on our Colibri Core research.

Keywords: Colibri Core, Translation Patterns, Syriac

1. Introduction

To what extent can linguistically uninformed features help us in tracing divergent patterns in an ancient Syriac Bible translation (the Peshitta, 2nd cent. AD) and its Hebrew source text? To answer this question, we need a language-independent tool that allows for a fine-grained comparison of both texts, without the need of textual annotations. The CLARIAH component Colibri Core, which has been applied in translation studies before, is very promising in this respect. This article summarizes the first results of the PaTraCoSy (PAtterns in TRAnslation: Using COlibri Core for the Hebrew Bible corpus and its SYriac translation) project, which is funded by CLARIAH. The overall goal of this project is to use the Colibri Core environment to detect translation patterns between the Hebrew Bible and the Peshitta, its Syriac translation, thus providing a follow-up of the CLARIAH research pilot Linking Syriac Data. The richly annotated linguistic database of the Hebrew Bible has been created over a period of almost four decades (1977–2017). Thanks to a CLARIN-NL project (2013–2014), it has been made available through the SHEBANQ website, besides its presence on GitHub as the BHSA. An electronic representation of the ancient Syriac translation of this text, called the Peshitta is also produced and maintained by the Eep Talstra Centre for Bible and Computer (ETCBC). Modifications of this corpus were made in the CLARIAH research pilot Linking Syriac Data (2017–2018). Some encoded texts (Kings, Psalms 1–30 and others) are linguistically annotated in a way similar to the Biblia Hebraica Stuttgartensia Amstelodamensis, the Hebrew Bible project of the ETCBC.

Colibri Core is a CLARIAH tool developed by van Gompel (2016) within the scope of his PhD research project. In van Gompel's dissertation, ColibriCore is used in automatic translation and word sense disambiguation based on context-sensitive suggestions for translations from one language into another. This is an interesting case in relation to the Bible, because Bible Translation and Machine Translation have been allies that need each other and reinforce each other: the Bible providing a huge parallel corpus in a few thousand language for Machine Translation, Machine Translation being the most advanced means to support and speed up Bible translation projects, as discussed for example by Hurskainen (2020).

The computation of n-gram pattern models - with the extension of skipgrams and flexgrams as provided by Colibri Core, where we skip a fixed amount or a flexible amount of words respectively - determines a basis for comparative corpus analysis. Since n-grams are typically distributed in a Zipfian fashion, there are only a few high-frequency patterns, with words such as common function words in the lead, and there is a long tail of patterns that occur only sparsely. Whether the same holds true for skipgrams and flexgrams remains to be determined, but a highly similar pattern is to be expected since they derive from regular n-grams. N-grams that are not subsumed by higher order n-grams, i.e., which do not occur as part of a higher order n-gram in the data/model, can be pruned from the model. This pruning allows us to focus on the most salient n-gram features. We use the Hebrew data as the baseline and compare to what extent these features persist in the Syriac translation. An important metric for corpus comparison is log-likelihood, which we will not yet discuss in this preliminary report. This metric expresses how

much more likely any given pattern is for either of the two models, which therefore allows us to identify how indicative a pattern is for a particular corpus.

In this project, we want to experiment from the opposite starting point of Colibri Core. We do not use it as a tool to create translations, but to discuss and compare existing translations. Using the n-gram, skipgram and flexgram search capabilities of ColibriCore, we can track significant word groups and their translations in parallel. In the PaTraCoSy proejct, we are mainly interested in two questions. 1. Do highly divergent translation patterns reflect specific syntactic differences? This question can be answered by comparing the ColibriCore output to annotations in the ETCBC database. 2. Which higher order n-gram receives a significantly other translation than its constituent parts? Answers to this question will describe patterns found in the translation of fixed expressions and other non-compositional structures. This is exactly why the ancient Syriac Bible translation provides an interesting text case, because Hebrew and Syriac are cognate languages, and yet have each their own structure. These questions are too broad to be answered within the confines of this article. For this reason, we limit ourselves firstly to a discussion of the word and phrase alignment process, a necessary step in order to use Colibri Core, and secondly to a description of first insights into the patterns found by the n-gram/skipgram analysis. These will be discussed in the following two sections, after which we formulate conclusions and ideas for future work.

2. Preliminary Work: Word and Phrase Alignment

Before we start our discussion of the alignment procedure, we need to determine the source text for the PaTraCoSy project. The ETCBC possesses a wide variety of textual encodings of both the Hebrew Bible and the Peshitta, leaving us a range of texts to choose from. It would lead us too far to discuss this variety in detail and explain why we chose our specific text. The variety ranges from text in the original Hebrew/Syriac (Syriac has three standard scripts) to transcriptions into Latin characters, over texts with or without vocalization and diacritical marks. Due to the richness of the annotations made through the texts over the course of many years and projects, all attested words are also accompanied by syntactic, morphological, semantic and prosodic information, determining very precisely which structures are attested throughout the texts. We chose a lemmatized text, stripped of vowel signs and other diacritics, so that we can focus on the structure of the use of lexemes in both languages.

2.1. Establishing the Alignments

The results of the preliminary work we discuss here can be found in our GitHub repository[1]. In the directory 'Genesis Alignment', all files necessary to find word and phrase alignments between the Hebrew and Syriac books of Genesis can be found. We focus on a specific book so that we can compare patterns between this book and the entire Bible in the pattern modelling part of this article. We choose the book of Genesis specifically, because it has received much attention from translation studies between Hebrew and Syriac, for example Morrison et al. (2019).

The word and phrase alignments can be found among the .txt files, where 'actual.ti.final' contains the word alignment, and 'AA3.final' the phrase alignment. In order to perform word and phrase alignment according to respectively Och and Ney (2003) and Koehn (2009), we implement the Giza++ library from the Moses-SMT Github page[2]. Using the plain2snt.out command, we first created vocabulary files for both Hebrew and Syriac, based on the input texts from the book of Genesis. The .vcb extension indicates these files. Once we have these files, we can use them in the snt2cooc.out command to generate a co-occurrence file, which is given the extension .cooc. This file in turn then is necessary to use the main Giza++ command, in order to construct the alignment files.

2.2. Basic Alignment Test

As we have described above, we assume Hebrew to be the source language, and Syriac the target. The file 'actual.ti.final' describes the word alignment, where the first word is the Syriac target, followed by a Hebrew source word and the alignment score based on the Viterbi algorithm. The file 'AA3.final' contains the sentence alignments, which consists of three lines of information. The first contains the alignment score, once again based on the Viterbi algorithm. The second line contains the Syriac target sentence, and the third the Hebrew source sentence. This final line always starts with NULL(), where words that did not receive an alignment are placed in between the curly brackets. This means that whenever no word is found here, all words are aligned to a target word or word group. An easy example can immediately be found in Genesis 1,1, where every source word stands in a one-to-one correspondence to a target word. Of course, most sentences are not that easy to align. In Genesis 1,2, for example, we find WXCK, aligned to WXCWK¿ ¡L, while ¡L should be aligned to the (first part of) ¡L PNJ.

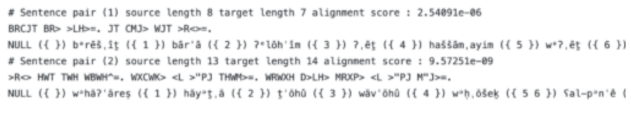

Figure 1: Sentence Pair 1

In this case, we can see that the preposition is not aligned correctly, but that words with semantic information still are correctly aligned. However, most

[1] https://github.com/ETCBC/PaTraCoSy

[2] https://github.com/moses-smt/giza-pp

of these prepositional inconsistencies between Hebrew and Syriac are correctly aligned. For example for sentence pair 320, where the Hebrew MMYRJM (one unit consisting of a preposition and substantive) is aligned correctly to the Syriac MN MYRJN= , which equally consists of a preposition and substantive, but this time separated into two words.

```
# Sentence pair (320) source length 10 target length 12 alignment score : 8.89016e-10
WSL#Q >BRM MN MYRJN=. H#W W>NTTH WKL D>JT LH=. WLWV <MH LTJMN>=.
NULL ({ }) wayyaʕal ({ 1 }) ʔavr̩ām ({ 2 }) mimmiṣr'ayim ({ 3 4 }) hû ({ 5 }) wᵊʔišt'ô ({ 6 })
```

Figure 2: Sentence Pair 2

This method can also be used to trace interesting differences between the Syriac translation and the Hebrew original, which we will use in our further research involving Colibri Core. For example sentence pair 369, where the Hebrew JHWH is linked to two Syriac words, ¿BRM MRJ¿. Only the second alignment is correct, where the word ¿BRM in Syriac does not have a source word in Hebrew. This is an example where the Peshitta mentions Abraham explicitly, but the Hebrew does not.

```
# Sentence pair (369) source length 8 target length 9 alignment score : 3.02109e-07
W>M#R >BRM MRJ> >LH>=. BMN> >D< DJRT^ >N> LH^=.
NULL ({ }) wayyōmʾar ({ 1 }) ʔᵃḡōn'āy ({ 2 3 }) [yᵉhw'ih] ({ 4 }) bamm̩ā ({ 5 }) ʔēḏ̩aʕ
```

Figure 3: Sentence Pair 3

Now that we have created the word and sentence alignments, we are ready to delve deeper into the main purpose of Colibri Core, namely pattern modelling.

3. Patterns in Translation

This third section finally brings us to a first discussion of the translation patterns from the Hebrew Bible into the Peshitta. Before we can address the results, we briefly describe how we constructed the model. After installation of Colibri Core, we follow the main thread in constructing the pattern modeling files, using the Colibri Core pattern modeller. We do not need tokenization of the texts, because of our very specific choice of base text, provided by the ETCBC. Our first step then is class encoding the data of the two text files, resulting in .cls and .dat files.

3.1. General Results

Colibri-Core allows us to trace recurring patterns in a monolingual corpus, not allowing a direct comparison in a bilingual corpus. For this reason, we make two models, one for Hebrew, one for Syriac, while counting the amount of constructions, be they n-grams or skipgrams. At the moment, we do not yet use the flexgram functionality of Colibri Core, because we need the log-likelihood computation to interpret these results. This did not fit into the confines of this article. We did not determine a maximum amount for the n-gram and skipgrams, only a base requirement of 10 attestations. This means that a word has to appear in at least 10 different constructions in order to be considered part of the model. We prefer indexed models, to be able to exactly trace back all patterns and their attestations within the corpus, although the memory requirements increase with this preference. We divided the corpus into newline segments for every new Bible verse, so the size of the skip gram will never exceed these boundaries, avoiding non-sensical combinations.
From the Colibri Core architecture, we can derive not only the type of data to be investigated, but also a succinct presentation in the form of reports and histograms.

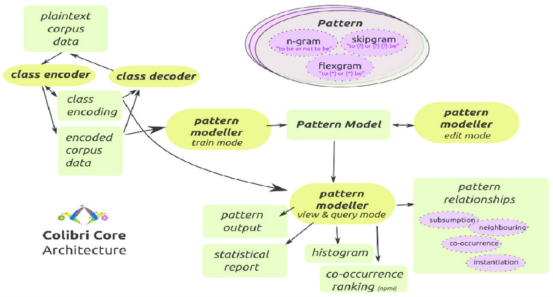

Figure 4: Colibri Core Architecture

The histogram of the entire biblical corpus for both languages can help us in detecting that Syriac and Hebrew have similar patterns, all within ten percent variance from each other. This also determines that, at least for this corpus, skipgrams too behave in a Zipfian fashion. The report shows us more interesting information. For most types of structures, for example n-grams for n=3 or skipgrams with maximum of two skips, we find that Syriac has more patterns, and correspondingly a smaller coverage in the entire corpus. This trend is nearly universal, leaving us to conclude that the Syriac translation is relatively flexible in that it provides several constructions corresponding to the Hebrew original. This overall trend can be applied to individual books, allowing us to deduce that Genesis has the widest variety of Syriac constructions compared to the Hebrew original, if we do not count poetic books, which are a separate category in dealing with translation patterns. This is another important reason to consider the book of Genesis as a first indicator of the types of translation patterns we can expect. In further research, we will be able to use these metrics to determine the typicality of certain translation patterns within the framework of the book in which they appear. This can be of great help to translation studies of ancient texts, since these patterns may reflect different authors, genres, and scribal schools.

3.2. Specific Translation Patterns

Apart from this general insight, it is difficult to derive more insights into the translations as a whole, without looking at specific translation choices first, and their

comparative place among other choices and patterns. A full description of these structures will allow us to discern general directions the translators took with respect to specific books, genres or semantic fields.

A first pattern we find is that where the translator chooses a word with a strongly different statistical distribution than its original. In general, the translator is not always consistent in his lexical choices, leading us to consider them term by term. In Genesis 2:6, for example, the Hebrew hapax legomenon (a word that occurs only once in the corpus) ¿D is translated by the more common Syriac word MBW¡¿. In this case, we cannot perform a comparative study of these terms, since there is only one single construction for Hebrew in which this term occurs. Sometimes, the Peshiṭta simply imports the Hebrew text into Syriac, leaving the often difficult to interpret or obscure meaning of the Hebrew undetermined. This in turn then produces hapax legomena in Syriac, leaving the interpretation fully to exegetes, who specialize in the interpretation of the original Hebrew text. In this sense, the Peshitta sometimes forsakes in its task of being a true translation, placing loyality to the original above ease of interpretation for the Syriac reader. When the translator imitates the Hebrew text, we are left guessing whether the translator understood the text in front of him.

A clear example of a specific translator choice being unloyal to the source text can be found in Genesis 2:2. The Hebrew BJWM HCBJ¡J 'on the seventh day' is translated into the Syriac BJWM¿ CTJTJ¿ 'on the sixth day'. The translation of numbers diverges only in this one single instance. This proves that this is a very exceptional situation, for theological reasons, rather than linguistic or cultural ones. On the other hand, the Hebrew BJWM occurs in 524 constructions, with a coverage 0.00171483, whereas in Syriac, this word occurs in 787 constructions with coverage 0.00249515. This allows us to conclude that Syriac uses the word for 'day' more freely than the Hebrew original.

A more specific example of the Syriac translation providing a concrete interpretation of the original can be found in Genesis 8:21. The Hebrew ¿T RJX HNJXX 'the pleasing odor' is translated into the Syriac RJX¿ DSWT¿ RJX¿ DNJX¿ 'the sweet fragrance, the fragrance of repose'. In Hebrew, RJX occurs in 30 constructions with coverage 0,0000981775. In Syriac, it only occurs in 16 constructions, with coverage 0,0000507273. This overlap is very strong, with this example being the only example of a non-literal translation for this word. The opposite process can be found in Genesis 9:14, where the Hebrew WHJH B¡NNJ ¡NN 'when my clouding with clouds' is rendered into a more straightforward Syriac DMSQ ¿N¿ ¡N"N¿'when I bring up clouds'. We learn that in Hebrew, ¡NN occurs in 43 n-gram constructions with a coverage of 0.0001407211, whereas the Syriac in 72, with a coverage of 0.000228273. Again, this is an example where the corresponding words in both language have nearly the same behaviour in n-gram/skipgram structure, except for a very specific case.

The final type of translation pattern is where a large variance between attested n-gram/skipgrams can be detected for semantically similar words, due to word play. A clear example can be found in Genesis 21:6. Here, the Hebrew YXQ... JYXQ LJ 'laughter... let him laugh with me' is translated into the Syriac XDWT¿... NXD¿ LJ 'great joy... let him rejoice with me'. The Hebrew YXQ occurs in 82 constructions, with coverage 0.000268352, while the Syriac equivalent occurs 935 times, coverage 0.00296438. The Syriac translation occurs more then ten times as often as its Hebrew original. Investigating more closely why this divergence might be so elaborate, we conclude that the Hebrew verb refers through wordplay to Isaac, who appears in this context and whose name is near the same as the verb. Since this verb does not translate using the same lexeme in Syriac, this wordplay cannot be rendered in Syriac, leaving the translator no choice than to use another, far more common, word.

4. Conclusions and Future Work

In this article, we have discussed how Colibri Core can be used to detect translation patterns between Hebrew and Syriac. As we have indicated, the results in this paper still reflect a work in progress. Future work will focus on several aspects. Firstly, we will reflect further on the detected patterns, in order to determine which translation patterns we want to discuss using the Colibri Core approach. This article has already provided several examples of structures which we should investigate more broadly over the entire Bible, rather than solely in the book of Genesis. In this stage of the research, we also propose our findings to a group of Syriac scholars, to further reflect on the important patterns to discuss further and to inquire for a more quantitative approach than they have hitherto received. We also provide a full phrase-translation table which, after some filtering, will be made accessible for feedback to Syriac scholars with the Colloquery web application. Earlier research by Van Peursen (2007) suggested that there are some recurring corresponding phrase patterns, e.g.: noun + abstract noun in Hebrew ("sons of wickedness") corresponding to noun + adjective in Syriac ("wicked men"), but his research, although "computer-assisted", largely remained within the boundaries of more traditional linguistic text comparison. It is to be expected that the linguistically uninformed techniques will bring to light other, unexpected patterns. Next to this more manual approach, we also expand our quantitative investigations, by calculating the log-likelihood values for the n-gram/skipgram patterns. This will then also allow us to consider flexgrams. We will not only consider the log-likelihood computations of these Colibri Core patterns, but also based on the translation patterns that arise from these structures. We hope that our research can contribute to a specialised debate between tradi-

tional textual scholarship and digital technologies.

Bibliographical References

Hurskainen, A. (2020). Can machine translation assist in Bible translation? Technical Reports on Language Technology Report 62.

Morrison, C.E., Kiraz, G., and Bali, J. (2019). The Syriac Peshitta Bible with English Translation. Genesis. Piscataway, NJ: Gorgias Press.

Koehn, P. (2009). Statistical Machine Translation. Cambridge University Press.

Och, F.J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics 29(1), pp. 19-51.

van Gompel, M. and van den Bosch, A. (2016). Efficient N-gram, Skipgram and Flexgram Modelling with Colibri Core. Journal of Open Research Software, 4(1), pp. 30-40.

van Peursen, W.T. (2007). Language and Interpretation in the Syriac Text of Ben Sira. A Compara-tive Linguistic and Literary Study (Monographs of the Peshitta Institute Leiden 16;). Leiden: Brill.