Proceedings of the 16th Conference on Computer Science and Intelligence Systems pp. 111–114

DOI: 10.15439/2021F30
ISSN 2300-5963 ACSIS, Vol. 25

Automated creation of parallel Bible corpora with cross-lingual semantic concordance

Jens Dörpinghaus*, Carsten Düing†
* University of Pretoria, Faculty of Theology and Religion, Hatfield, Pretoria, South Africa,
Email: u21829927@tuks.co.za
† Carsten Düing has been with the Faculty for Mathematics and Informatics,
Fernuniversität Hagen, Germany at the time of this work

***Abstract*—Here we present a novel approach for automated creation of parallel New Testament corpora with cross-lingual semantic concordance based on Strong's numbers. As scientific editions and translations of Bible texts are often not free to use for scientific purposes and are rarely free to use, and due to the fact that the annotation, curation and quality control of alignments between these texts are quite expensive, there is a lack of available Biblical resources for scholars. We present two approaches to tackle the problem, a dictionary-based approach and a Conditional Random Field (CRF) model and a detailed evaluation on annotated and non-annotated translations. We discuss a proof-of-concept based on English and German New Testament translations. The results presented in this paper are novel and according to our knowledge unique. They present promising performance, although further research is necessary.**

I. Introduction

Building a concordance of texts, automated text alignment and automated text translations are well studied topics in research. A *semantic concordance* is a widely used approach to link text corpora with data and values in lexicons, see [1]. Even in the humanities a lot of research has been done within this wide field of text mining and automated text processing. Coming to the field of what some call *Digital Theology* as a subfield of *Digital Humanities* and its intersection to ancient languages we still see a lot of challenges, although the problems themselves may seem easy and a standard task.

Here, we want to tackle the challenge of automated annotations of words within New Testament texts to create parallel Bible corpora in different languages. Thus, our goal is to create a cross-lingual concordance alignment for New Testament texts and translations. These are widely used for research and teaching. Our approach is restricted to the mapping between original and translated words given both the translation with or without further information and the Greek source with morphological information.

Research on Biblical texts and translations of course has a long tradition and translations have been widely used.There was a great increase in the amount of different Bible translations in the nineteenth century and thus, the research in this field increased also, see [2]

New approaches from computer science have also been used to evaluate translations and texts but only really took off in the last 30 years as it became more accessible to scholars with a different background. It is possible to use these methods to understand the manual curation and understanding of text and it would be to improve the technological solutions for automated approaches. Here, Clivaz states in 2017 [3] that only very little research has been done in this field and Anderson underlines the theologians lack of interest for digital and modern text mining methods a year later [4]Only the fields of digital manuscripts, Digital Academic Research and Publishing show some progress [5]. This work tries to be a first step to close this gap.

As scientific editions and translations of Bible texts are often not free to use and due to the fact that working on them is quite expensive, there is a lack of available Biblical resources for scholars. The aim of this work is to develop and evaluate novel approaches for automated generation of alignment for parallel Bibles leading to cross-lingual semantic concordance.

This paper is divided into six sections. The first introduces the problem. The second section gives a brief overview over the state of the art and related work. The third section is dedicated to the data foundation. We will also discuss the annotation style and the selection of training and test data. In the fourth section, we present two approaches to tackle the problem. We introduce a dictionary-based approach and a CRF model. The fifth section is dedicated to experimental results on annotated and non-annotated translations. Our conclusions are drawn in the final section. The results presented in this paper are novel and according to our knowledge unique. They present promising performance, although further research is necessary.

II. Related Work

Since only little research has been done in this field, we list all material available even if there tasks are only tangentially related. In Biblical research *The Exhaustive Concordance of the Bible* from 1890 is widely used to link words from Biblical texts to dictionary entries. These so-called Strong's numbers can be used to create automatic aligned parallel texts, see [6] or [7] who created semantic maps from parallel text data. Here, texts in multiple languages are presented together [8]. Although a lot of approaches are based on machine translation in Biblical research, these texts are still mainly hand-crafted, e.g. [9] or [10].Even if the Bible is often used as training model or reference model for unsupervised learning models for translation, see for example [11], [12] or [13], only few approaches have also been made to analyze religious or theological texts with methods from AI and text mining.

To cover the language related question other scholars examined the impact of computer technologies on Bible translations and discussed their limitations [14]. Bible translations usually not being in the scope of linguistic research, but interesting for the history of language, there is a wide range of publications and analyses of recent translations, see e.g. [15] and [16]. There is also a considerable amount of literature on Bible translations [17]. It is important to notice that Bible translation is not only about decisions between translation strategies like formal or dynamic equivalence.

Encoding linguistic information in multi-language documents produces *Interlinear Glossed Text* (IGT). Biblical texts are usually well-studied and thus both references to the Strong's numbers as well as morphological information are available for Hebrew and Greek texts. Automated glossing is also a widely studied field, see [18] or [19]. These approaches have never been used to create interlinear glossed Biblical texts. Only some little research has been done on the Qur'an [20]. For automated translations, there are no resources available for ancient Greek [21]. Other approaches, like GASC [22], build a Bayesian model describing evolution of words and meanings in ancient texts. They state "a lack of previous works that focussed on ancient languages". Thus, not only the target texts form a new field, but we can also only build upon very little work within the field of automated translations.

III. Data

Here, we will focus on Greek original text, German and English Bible translations, although this approach can be used for any other language. There are several software packages available to access Biblical texts. Some commercial software like Logos offer no or only very limited access to their API[1]. Thus, we did our work on the basis of the SWORD Project, which offers a full API available under GNU license[2]. As a basis for the Greek text we used the SBLGNT 2.0 from Tyndale House, based on SBLGNT v.1.3 from Crosswire. This text is with some minor changes comparable to the Nestle-Aland/United Bible Societies text. The English texts are based on KJV (1769, King James Version), ASV (1901 American Standard Version) and ESV (English Standard Version, 2011). The dictionaries are based on the original Strong's Dictionaries or are extracted from the texts. The German texts are based on Luther (1912), Leonberger Bibel (2017), the Greek-German dictionary by Gerhard Kautz and for a detailed analysis on some excerpts of newer translations. Beside of them, all data is available with a free license. See http://www.crosswire.org/sword/modules/ for details of these packages.

Different approaches for translating Biblical texts exists. KJV, ESV and ASV follow a traditional word-for-word approach, also known as formal equivalence. The Leonberger Bibel follows the same approach, whereas Luther 1912 also has elements from the thought-for-thought approach known as dynamic equivalence. For testing purposes we will also consider translations which use a paraphrase approach. For a detailed overview about Bible translations see [2].

A. Annotation Style

There are several annotations which can be displayed in different ways. Here, we rely on the HTML-output. Both lemma and morphology information are included in w-tags. For example in Acts 1:1:

```
<w lemma="strong:G3303" morph="
    robinson:PRT" savlm="strong:G3303" src
    ="2"/>
```

We will use this annotation style both for extracting information, storing and comparing them.

B. Training and Test Data

To collect the training data, we can use the complete New Testament texts mentioned above. This leads to 7,957 verses in each version. There are 5,624 entries in the Strong's dictionary. We tested our models both on a random subset of the same and other translations. In addition, we will test our model on some verses from newer versions, e.g. the recent German Luther-Bible. Here, the verses are curated by hand.

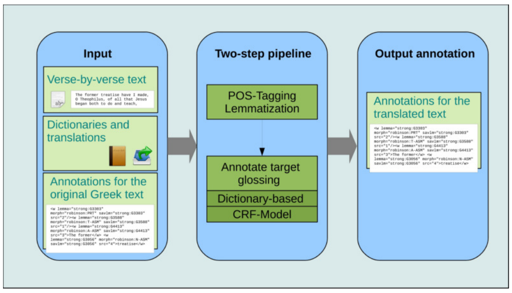
IV. Methodology

A. Modeling

Here, we have Biblical texts containing verses. Each verse X contains a sequence of words, thus $X = x_1, ..., x_N$. Given two languages L and L' we have two sequences

$$\begin{aligned} X^L &= x_1^L, ..., x_N^L \\ X^{L'} &= x_1^{L'}, ..., x_M^{L'} \end{aligned}$$

And we want to model the target glossing $f : X^L \to X^{L'}$ that contains mappings from a word origin $x_i^L \in X^L$ to another word $x_j^{L'} \in X^{L'}$. Let Y be a sequence of all mappings, than we need to compute $P(Y|X^L)$.


Input
Verse-by-verse text
Dictionaries and translations
Annotations for the original Greek text
Two-step pipeline
POS-Tagging Lemmatization
Annotate target glossing
Dictionary-based
CRF-Model
Output annotation
Annotations for the translated text

```
The former treatise have I made,
O Theophilus, of all that Jesus
began both to do and teach,
```

```
<w lemma="strong:G3303"
morph="robinson:PRT" savlm="strong:G3303"
src="2"/><w lemma="strong:G3588"
morph="robinson:T-ASM" savlm="strong:G3588"
src="1"/><w lemma="strong:G4413"
morph="robinson:A-ASM" savlm="strong:G4413"
src="3">The former</w> <w
lemma="strong:G3056" morph="robinson:N-ASM"
savlm="strong:G3056" src="4">treatise</w>
```

```
<w lemma="strong:G3303"
morph="robinson:PRT" savlm="strong:G3303"
src="2"/><w lemma="strong:G3588"
morph="robinson:T-ASM" savlm="strong:G3588"
src="1"/><w lemma="strong:G4413"
morph="robinson:A-ASM" savlm="strong:G4413"
src="3">The former</w> <w
lemma="strong:G3056" morph="robinson:N-ASM"
savlm="strong:G3056" src="4">treatise</w>
```

Figure 1. The proposed two-step method. First, we use a POS-Tagging and Lemmatization to extract the word to be matched. Then, we annotate the target glossing, either with the dictionary-based method or using a CRF-Model. As input, we use the target text (a translated text) verse-by-verse, the original Greek text containing the annotations and some additional information from dictionaries and Biblical translations.

[1] See for example https://wiki.logos.com/Logos_4_COM_API.

[2] See http://crosswire.org/sword/index.jsp

Algorithm 1 DICTIONARY-BASED-MATCHES

```
Require: Sequences of words $X^L = x_1^L, ..., x_N^L$ with dictionary mapping to $d(x_i^L)$ to dictionary $D$ and in target language $X^{L'} = x_1^{L'}, ..., x_M^{L'}$.
Ensure: Mapping $f : X^L \rightarrow X^{L'}$.
    for $c$ in $POS$: do
        for $x_i^L$ in $c$: do
            find $x_j^{L'}$ with $\min \delta(lem(d(x_i^L)), lem(x_j^{L'}))$
            assign $f(x_i^L) = x_j^{L'}$
        end for
    end for
    return $f$
```

Here, we propose a two-step method. As input we use the target text (a translated text) verse-by-verse, if needed, the original Greek text containing the Strong's annotations and some additional information from dictionaries and Biblical translations. First, we use POS-tagging and lemmatization to extract the word to be matched. Then, we annotate the target glossing, either with the dictionary-based method which is a natural fit because a lot of features are available or using a CRF-Model which is one of the standard solutions in current NLP. See figure 1 for an illustration.

B. *Dictionary-based approach*

After detecting parts-of-speech in the target text, we can sort words from the original Greek text and the target language. This helps to reduce the target set of words. Since we know the Greek Strong's numbers, we can use lemmatization to compare words and assign the best fit, see algorithms 1.

We need to choose a proper distance function δ (like Levenshtein distance or cosine similarity) and we need to choose proper dictionaries.

By language, we can either rely on dictionaries: the Greek-English dictionary by Dr. Ulrik Sandborg-Petersen and the Greek-German dictionary by Gerhard Kautz, both released under CC license. In addition, we build dictionaries from the annotated Biblical texts presented in section III. Here, we for every Strong's number we collected words in a particular Bible translation.

In order to make this data available, we wrote an importer to create a list of words in the target language associated with a Strong's number. This dictionary-based approach is a lazy learner approach, since first we learn dictionaries but the comparison and assignment is done in a separate step. Thus, we will now introduce a different approach using CRF models.

C. *CRF-Model*

Our second approach uses a linear-chain Conditional Random Field (CRF), see [23]. Here, we train a sequence model where the input consists of words and the output of a Strong's-labels. [18] used this method to automatise the gloss generation in interlinear glossed texts. Here, we used sklearn-crfsuite v0.3.6 to build the CRF models. For training, we used stochastic gradient descent with L2 regularization and a maximum of 50 iterations.

Target	Luther1912			GerLeoNA28		
Source	P	R	F1	P	R	F1
Luther1912	.75	1	.84	.45	.83	.55
GerLeoNA28	.69	.96	.78	.56	.95	.67
Luther1912 + GerLeoNA28	.77	1	.86	.57	.92	.67
CRF Luther1912	.12	.14	.13	-	-	-
CRF GerLeoNA28	-	-	-	.53	.52	.52

Table I
EVALUATION OF DIFFERENT GERMAN TARGET TRANSLATIONS. THE BASIS ARE EITHER A COMBINATION OF DICTIONARY-BASED APPROACHES OR THE CRF MODEL. HERE, P REFERS TO PRECISION, R TO RECALL, F1 TO F1-SCORE.

Target	KJV			ASV			ESV		
Source	P	R	F1	P	R	F1	P	R	F1
KJV	.50	.83	.58	.69	1	0.79	.63	.96	.74
ASV	.44	.79	.53	.73	.96	.81	.66	.96	.75
ESV	.38	.71	.46	.72	.96	.80	.67	.96	.78
ASV + ESV	.38	.71	.46	.72	.96	.80	.68	.96	.78
KJV + ASV + ESV	.41	.75	.49	.71	.96	.79	.68	.96	.78
CRF KJV	.26	.20	.20	-	-	-	-	-	-
CRF ASV	-	-	-	.33	.33	.33	-	-	-
CRF ESV	-	-	-	-	-	-	.27	.25	.25

Table II
EVALUATION OF DIFFERENT ENGLISH TARGET TRANSLATIONS. THE BASIS ARE EITHER A COMBINATION OF DICTIONARY-BASED APPROACHES OR THE CRF MODEL. HERE, P REFERS TO PRECISION, R TO RECALL, F1 TO F1-SCORE.

For our testing purpose, we include basic linguistic features, the source and previous and following words. For training purposes, we can again rely on the Biblical texts which are already annotated with Strong's numbers.

D. *Evaluation*

The performance of each approach is evaluated by comparing each annotation in each final output to the test data set provided from annotated Biblical texts. Thus, we need to cross-evaluate different input scenarios against different and similar output scenarios. The Greek-English dictionary by Dr. Ulrik Sandborg-Petersen and the Greek-German dictionary by Gerhard Kautz were not presented, because it was not possible to extract the exact proposed translations with reasonable effort.

Since our approach produces Strong's numbers annotations for words in translated text, the first question is if this leads to proper assignments on the *same* text. We will also evaluate, if combining different models will lead to better solutions. Because these approaches may predict Strong's numbers that have more or fewer occurrences in the text we add both precision and recall to our evaluation. These metrics are presented as a micro-average value over all verses.

Further, we will analyze how these systems will work on unanotated translations. For this purpose, a few verses have been chosen to evaluate the output.

V. RESULTS

A detailed evaluation with to precision, recall, and F1-Score can be found in tables I for German translations and II for

English translations. These tests showed unexpectedly that the CRF models were not competitive with the dictionary-based approaches. We will discuss this observation and possible reasons later.

The dictionary based approaches on German translations (table I) show very promising results. The recall value is really high, although the precision value increases, if more dictionaries are combined. Although we can see a different behavior for Luther1912 and GerLeoNA28. For the latter, the combination of both dictionaries increases the precision, but also decreases recall value. It is crucial to note that a combination of dictionaries needs a careful investigation. Here, one of the reasons might be that although both translation have been done with the same approach, there is more than hundred years in between them. So the words and their meanings might have changed. In the next section, we will do some preliminary observation on more recent translations.

This is even more significant for the evaluation of English translations in table II. ESV and ASV are both bases on KJV and again there are more than hundred years in between (1769, 1901, 2011). The two most recent translations show a good result, the recall value is high and the precision value increases with the matching dictionary. The most remarkable can be found when using KJV for a combination of dictionaries, it decreases the values significantly. This result has further strengthened our confidence that it is crucial to evaluate the dictionary basis for this approach.

VI. Conclusion and Future Work

This paper has described a first approach to automated annotations of words within New Testament texts to create parallel bible corpora in different languages to create a cross-lingual concordance alignment for New Testament texts and translations. We proposed a lazy-learner and an eager-learner approach: A dictionary-based and a CRF-based approach.

While the amount of training data was due to strict license politics in the field of Theology relatively low, we could nevertheless get promising results for some translations. This method can't be applied to translations following a paraphrase approach. This will hopefully lead to further research and a better understanding of special requirements within the field of theology and in particular ancient languages. Here, we see the need for more models and methods since there are no resources available for ancient Greek.

Our analysis of errors reveals a number of questions and also possible further improvement. First, we need to consider if more translations and Biblical texts can be used as training data. Although not in every case the results could be improved when more dictionaries were used a better data foundation together with improvements in modeling and algorithms will improve the results. Second, we need to investigate why some parts of speech, in particular nouns and conjunctions, do not work well at all. Finally, we need to make an in-depth error analysis why the CRF models do not work as expected. Here, we will invest weather a better feature selection (for example POS tagging or dependency labels) will improve the results.

While our proof of concept is both working and generic it is still very early work on a problem which needs more attention. We hope that it will also highlight the importance of more interdisciplinary research in this field.

References

[1] S. Landes, C. Leacock, and R. I. Tengi, "Building semantic concordances," *WordNet: An electronic lexical database*, vol. 199, no. 216, pp. 199–216, 1998.

[2] B. Metzger, *The Bible in Translation: Ancient and English Versions*, ser. Biblical studies. Baker Publishing Group, 2001.

[3] C. Clivaz, "Die bibel im digitalen zeitalter: Multimodale schriften in gemeinschaften," *Zeitschrift für Neues Testament*, vol. 20, no. 39/40, pp. 35–57, 2017.

[4] C. Anderson, "Digital humanities and the future of theology," 2018.

[5] C. Clivaz, A. Gregory, and D. Hamidović, *Digital Humanities in Biblical, Early Jewish and Early Christian Studies*. Brill, 2013.

[6] M. Cysouw, C. Biemann, and M. Ongyerth, "Using strong's numbers in the bible to test an automatic alignment of parallel texts," *STUF-language typology and universals*, vol. 60, no. 2, pp. 158–171, 2007.

[7] B. Wälchli, "Similarity semantics and building probabilistic semantic maps from parallel texts," *Linguistic Discovery*, vol. 8, no. 1, pp. 331–371, 2010.

[8] M. Simard, "Building and using parallel text for translation," *The Routledge Handbook of Translation and Technology*, pp. 78–90, 2020.

[9] A. Yli-Jyrä, J. Purhonen, M. Liljeqvist, A. Antturi, P. Nieminen, K. M. Räntilä, and V. Luoto, "Helfi: a hebrew-greek-finnish parallel bible corpus with cross-lingual morpheme alignment," *arXiv preprint arXiv:2003.07456*, 2020.

[10] N. Rees and J. Riding, "Automatic concordance creation for texts in any language," *Proceedings of Translation and the Computer*, vol. 31, 2009.

[11] M. Diab and S. Finch, "A statistical word-level translation model for comparable corpora," MARYLAND UNIV COLLEGE PARK INST FOR ADVANCED COMPUTER STUDIES, Tech. Rep., 2000.

[12] P. Resnik, M. B. Olsen, and M. Diab, "The bible as a parallel corpus: Annotating the 'book of 2000 tongues'," *Computers and the Humanities*, vol. 33, no. 1, pp. 129–153, 1999.

[13] C. Christodouloupoulos and M. Steedman, "A massively parallel corpus: the bible in 100 languages," *Language resources and evaluation*, vol. 49, no. 2, pp. 375–395, 2015.

[14] J. D. Riding, "Statistical glossing, language independent analysis in bible translation," *Translating and the Computer*, vol. 30, 2008.

[15] J. Renkema and C. van Wijk, "Converting the words of god: An experimental evaluation of stylistic choices in the new dutch bible translation," *Linguistica Antverpiensia, New Series–Themes in Translation Studies*, no. 1, 2002.

[16] L. De Vries, "Bible translation and primary orality," *The Bible Translator*, vol. 51, no. 1, pp. 101–114, 2000.

[17] G. G. Scorgie, M. L. Strauss, S. M. Voth *et al.*, *The challenge of Bible translation: Communicating God's Word to the world*. Zondervan Academic, 2009.

[18] A. McMillan-Major, "Automating gloss generation in interlinear glossed text," *Proceedings of the Society for Computation in Linguistics*, vol. 3, no. 1, pp. 338–349, 2020.

[19] X. Zhao, S. Ozaki, A. Anastasopoulos, G. Neubig, and L. Levin, "Automatic interlinear glossing for under-resourced languages leveraging translations," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 5397–5408.

[20] A. B. Muhammad, *Annotation of conceptual co-reference and text mining the Qur'an*. University of Leeds, 2012.

[21] E. Biagetti, C. Zanchi, and W. M. Short, "Toward the creation of wordnets for ancient indo-european languages," in *Proceedings of the 11th Global Wordnet Conference*, 2021, pp. 258–266.

[22] V. Perrone, M. Palma, S. Hengchen, A. Vatri, J. Q. Smith, and B. McGillivray, "GASC: Genre-aware semantic change for Ancient Greek," in *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 56–66. [Online]. Available: https://www.aclweb.org/anthology/W19-4707

[23] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," *Proceedings of the Eighteenth International Conferenceon Machine Learning*, 2001.