UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

The EDGeS Diachronic Bible Corpus

Bouma, G.; Coussé, E.; Dijkstra, T.; van der Sijs, N.

Publication date
2020
Document Version
Final published version
Published in
LREC 2020
License
CC BY-NC

Link to publication

Citation for published version (APA):
Bouma, G., Coussé, E., Dijkstra, T., & van der Sijs, N. (2020). The EDGeS Diachronic Bible Corpus. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *LREC 2020: Twelfth International Conference on Language Resources and Evaluation : May 11-16, 2020, Palais du Pharo, Marseille, France : conference proceedings* (pp. 5232-5239). The European Language Resources Association. https://aclanthology.org/2020.lrec-1.644

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:13 Nov 2026

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5232–5239
Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

The EDGeS Diachronic Bible Corpus

Gerlof Bouma*, Evie Coussé*, Trude Dijkstra×, Nicoline van der Sijs+
* University of Gothenburg, × University of Amsterdam, + Dutch Language Institute, Leiden
{gerlof.bouma, evie.cousse}@gu.se, g.w.h.dijkstra@uva.nl, nicoline.vandersijs@ivdnt.org

Abstract

We present the EDGeS Diachronic Bible Corpus: a diachronically and synchronically parallel corpus of Bible translations in Dutch, English, German and Swedish, with texts from the 14th century until today. It is compiled in the context of an intended longitudinal and contrastive study of complex verb constructions in Germanic. The paper discusses the corpus design principles, its selection of 36 Bibles, and the information and metadata encoded for the corpus texts. The EDGeS corpus will be available in two forms: the whole corpus will be accessible for researchers behind a login in the well-known OPUS search infrastructure, and the open subpart of the corpus will be available for download.

Keywords: Germanic languages, historical linguistics, parallel corpora

1. Introduction and background

We present the EDGeS Diachronic Bible Corpus, a synchronically and diachronically parallel corpus of Bible translations in English, Dutch, German and Swedish, spanning six and a half centuries.

The EDGeS corpus is constructed in the context of a diachronic study of *complex verb constructions* in Dutch, English, German and Swedish. Complex verb constructions are combinations of a main verb with multiple auxiliary verbs. The grammaticalization of auxiliary verbs in the Germanic languages is well-studied (Hilpert, 2011), and can already be observed in the earliest known sources. The rise of the possibility to combine multiple auxiliaries is not nearly as well-charted, but the earliest known attestations of combinations of modals in Dutch and English are from the 13th century (Coussé, 2015). In later stages of English, the double modal construction went into disuse, whereas it thrived and expanded in Dutch. Still, not much is known on why and how complex verb constructions came into being, and how they developed in the history of the Germanic languages.

In the context of this research, we have the following wish list for the corpus that will serve as our empirical base material: First, it should support studies with a longitudinal perspective. Including re-translations at regularly spaced intervals would facilitate such research. Secondly, the construction is not highly frequent, so a corpus of sizeable texts is needed. Thirdly and finally, we wish to compare the development of complex verb constructions between several Germanic languages, so we wish to include parallel translations from a time period in different languages. We thus need a diachronically and synchronically parallel corpus with a sufficient amount of corresponding material. The Bible is the only text that could form the basis of such a corpus: it is large enough (½–1M token range) and has been repeatedly translated into our languages of interest over a long time.

To give a small impression of the kind of data we expect to get from a parallel Bible corpus, consider the three-verb complex verb construction in 1 a (Dutch, 2004), and its aligned passages from two older Dutch Bibles as well as two English ones; all from 2 Kings 5:12. The sentence in 1 a combines a finite verb *had*, realizing past tense and irrealis mood, a non-finite modal *kunnen* 'be able to' and the main verb *baden* 'wash'. The corresponding clause in 1 b (Dutch, 1657) presents an earlier example of a complex verb construction, but note the use of another finite verb. Finding earlier attestations of subtypes of complex verb constructions can further our understanding of the temporal development of the construction, as can the differences in combinatorial possibilities and tendencies. Examples 1 c (Dutch, 1528) and 1 d (English, 1535) each only contain one verb, the former in the subjunctive, the latter inside an if-clause. Such correspondences let us study how similar meaning components can be expressed through inflection or clause type, or by periphrasis like in the complex verb construction.

(1) (a) Had ik me daarin niet kunnen baden [...]?
(b) soude ick my in die niet konnen wasschen, [...]?
(c) dat ick mi daer in wiesche, [...]?
(d) Yf J washe me also in them, [...]?
(e) May I not wash in them, [...]?

We refer to the project website[1] for future publications based on the corpus presented here. Instead, this paper focuses on the EDGeS corpus itself, and accompanies its release as resource for historical linguistic and computational linguistic research. We consider the design and compilation of the corpus, and how it will made available. The paper is structured as follows: Section 2 looks at existing parallel Bible corpora. Section 3 outlines the corpus design principles. The verse-level alignment is discussed in Section 4, and the corpus' availability in Section 5.

2. Related work

Many researchers before us have noted the potential of the Bible as a source for a parallel corpus. For one set of papers, the focus lies on the availability of a great number of different translations of this text. An early example in the

[1] spraakbanken.gu.se/en/projects/complex-verb-constructions

field of computational linguistics is the work reported in Resnik et al. (1999), on an aligned corpus of, at the time of publication, 8 Bibles and 27 New Testaments. The authors underline the advantages of the Bible as an 'widely available, representative sample of carefully translated texts in a variety of styles in a broad range of languages' (ibid., p129), and describe the process of automatic conversion of the different source formats into a unified SGML format. In a very similar spirit, Christodouloupoulos and Steedman (2015) present a corpus of 100 texts. Both these papers also discuss ways in which the material can be used. The massively parallel Bible corpus of Mayer and Cysouw (2014) contains over 900 texts from 840 languages. Originating in research on language typology (Cysouw and Wälchli, 2007), the focus in this corpus has been to collect translations into many different languages from many language families. The project is also unique in providing the Bible translations in a minimal, standardized format, that abstracts away from translation specific information such as the order of the Bible books, but that is more convenient computationally when dealing with many parallel texts than a deeply structured format like XML.

Other parallel Bible translation corpora have been much more modest in terms of the number of parallel texts – focusing instead on annotation and application. The PROIEL Treebank of ancient Indo-European languages contains aligned New Testament texts in 5 languages, with detailed morpho-syntactic, information structure and alignment information (Haug and Jøhndal, 2008). The Konstanz Resource of Questions has 4 Bible translations, and comes with (semi-)automatic annotation of two question types (Kalouli et al., 2018). The Biblia Medieval project (Enrique-Arias and Pueyo Mena, 2008–) offers diplomatic and normalized parallel transcriptions of 14 Spanish medieval and renaissance Bible translation manuscripts, with links to digital facsimile. The texts have also received part-of-speech annotations.

The clear diachronic and synchronic dimensions in the EDGeS corpus are at the heart of its design. This is also true of the corpus of Bible translations, fragments and related texts (harmonies, paraphrases) introduced in Chiarcos et al. (2014). This corpus contains just under 40M tokens in over 200 texts from 14 Germanic languages/language stages. In comparison, however, our corpus design follows much stricter criteria for which texts are to be included (see Section 3), resulting in a selection that is more balanced in size and time, and with texts that are parallel to a higher degree, albeit with fewer texts and from a smaller number of languages. The work reported by Breder Birkenes et al. (in press) is also based on data that is parallel in the synchronic and diachronic dimension. However, because of the potentially much higher incidence rate of the studied agreement phenomena, parallel texts consisting of just one chapter from the New Testament sufficed for their purposes. As mentioned in the introduction, to support the study of complex verb constructions, the collection of larger texts is crucial.

3. Selection and collection

3.1. Main principles

As sketched in the introduction, we set out to construct a corpus of sizeable, diachronically and synchronically parallel texts. Our main principles and preferences for selection of the Bibles can be listed as follows:

- We are looking for Dutch, English, German and Swedish translations,
- from 1300 until present day, with at least one Bible from each century.
- We prefer, but not limit ourselves too, translations
  - that are first editions of complete Bible translations (see remarks below), and not modernizations;
  - that are translations in a narrow sense (not: harmonies, paraphrases, rhyming Bibles, etc.) into a language variety that was current at the time of publication;
  - made a historical impact, typically through wide dissemination; and
  - whose text is available electronically, with a clear link to the original (but see below and the next section).

The Bible translations selected using these principles are listed in Table 1. Some discussion of these principles is warranted, since they sometimes conflict and sometimes are overruled by historic reality.

The wish to have a corpus with a high degree of parallelism is the reason to exclude translations in a wider sense (but see, on the one hand, de Vries, 2007, on why there is great variation between Bible translations, limiting parallelism between texts, and Chiarcos et al., 2014, on the other, on the possibilities of linking material that is not strictly parallel). Maximization of the size of each text underlies our focus on 'complete Bibles', by which we mean translations of both Testaments.[2] Note that the exact contents of such a complete Bible will nevertheless differ, depending on the used source text, what the translators considered to be part of the canon, etc. In a few cases we have chosen to include an incomplete Bible, to fill a gap in the table and/or to be able to include a translation of particular importance. Two examples are the oldest Dutch translation in our corpus, the *Hernse bijbel* (Dutch, 1361), which contains a restricted selection of Bible books, and the authoritative *Gustav Vasas bibel* (Swedish, 1541), of which we only have an incomplete electronic text, even though the printed text constitutes a complete Bible.

Since the translations vary in which material they include (the conception of canon varies, or sometimes non-canonical material is purposely included), increasing the size of the material and increasing the amount of parallelism in the material are opposing aims: one either includes any book

[2]This completeness refers to the translation itself. In a few cases, the Bibles we included are only partial, since they are not currently fully digitized. This can be addressed in future corpus updates.

that occurs in any translation (union) or only those books that appear in all translations (intersection). We opt for the first strategy and include as many books as available digitally for a given translation, even if this means including books that do not appear in all texts of the corpus. As a result, the corpus not only contains Old and New Testament books, but also a series of Apocryphal books.

Because we want to use the corpus to learn about a particular language stage, it is important to have translations that are representative of at least some variant of the language as used at the time of publication.[3] It is therefore desirable to avoid re-editions and (orthographic) modernizations, as they represent the language of the first edition in many linguistically relevant aspects. The same goes for translations that are intentionally archaic or try to mirror linguistic aspects of the source in the target language. Here, too, we may deviate to allow satisfying another preference. For instance, we have included the *Statenvertaling* (Dutch, 1657), which at the time of publication was already considered archaic, but ended up being the most influential Dutch translation for centuries to come.

The publication of a complete Bible may involve different editions of parts of the translation. For instance, *Challoner's Revision* (English, 1750/52) combines a first edition of his Old Testament with a third edition of his New Testament. In these cases, we have chosen the versions that constitute the complete Bible.

Electronic versions of the historical Dutch Bible translations were sourced from existing electronic editions made by the volunteer network *Stichting Vrijwilligersnetwerk Nederlandse Taal* (SVNT), under co-supervision of this paper's fourth author (Beelen and van der Sijs, 2014).[4] As part of the construction of the EDGeS corpus, the SVNT has created electronic versions of five additional New Testament translations, which has allowed us to fill in several gaps in the Dutch and German parts of the corpus.

As can be seen from Table 1, our goal of one translation per century is not quite met. In the case of the English 15th century, the gap is due to a ban on Bible translations in the vernacular (Wansbrough, 2008). For Swedish, we have included only four versions in total. The lack of historical translations in this part of our corpus is because of the dominance of *Gustav Vasas bibel* (1541): until the 19th century, nearly all complete Bible translations were re-editions and moderate revisions of the 1541 translation. Even the included *Karl XII:s bibel* (1703) is such a revision, although one that was very widely spread and in use under a long time (Dagson, 2013; Pettersson, 2017).

[3] It will, of course, always be the case that a text as specific as a single Bible translation will only give us a very narrow view of what 'a language' is like. In this short paper, we wish to focus more on the design and collection of the corpus than the methodological issues in using the resulting corpus as the basis for linguistic research. We refer to Kaiser (2005), de Vries (2007) and Rosemeyer and Enrique-Arias (2016) for some methodological discussion of using Bible translations.

[4] Also see the data archive at doi:10.17026/dans-xvx-frex

3.2. Metadata

The corpus has to be able to support historical linguistic research – this is, as described, the in-project goal of the corpus, and the wider historical linguistic community forms an important part of our intended audience. It is crucial that we give the user a clear idea to what extent they can trust the corpus. The metadata accompanying the materials therefore contain as much information as possible about the original Bible translation and about the provenance of the electronic transcription – whether we know who created it, whether there is a description of the digitization principles, etc. The transcriptions can be divided into three groups:

- manual diplomatic transcriptions from the SVNT network,
- third party transcriptions/digitizations with a clear explanation of the method – these also include copyrighted contemporary versions obtained from the translating bodies,
- third party versions with unknown/unclear digitization history – among these many that have been obtained from on-line sources with a focus on Bible study, such as `eBible.org` and `www.crosswire.org/sword`.

The inclusion of the last group presents a potential problem, as we do not know what the chain leading from the original to the electronic version looks like, and it therefore becomes harder to draw conclusions about the original by looking at the electronic version. We have tried to amend this situation by comparing (scans of) the printed editions to the available electronic text, so that we may judge to what extent the latter is a faithful transcription of the former. Even when we cannot definitively establish the source of an electronic version, this information is of value for the user, since they are then aware that further research is needed before conclusions can be made from the text as present in our collection.

4. Conversion and alignment

The texts come from a wide range of sources and in a myriad of formats with different levels of markup. The printed Bibles and manuscripts themselves differ in the extent to which they include paratexts such as comments, cross-references, divisions other than book-chapter-verse (stories, paragraphs, etc), and in addition, the electronic editions differ greatly in the extent to which they preserve this information and explicitly mark it as such. We are most interested in those parts of the Bibles that are likely to have counterparts in the other Bibles, so we primarily target the verses themselves for extraction. In addition, book and chapter titles are included, to give the researcher who accesses part of the material linearly (as opposed to through querying) a better frame of reference. Finally, introductory and concluding sentences can have varying status between the Bibles: they may be off-set typographically, unambiguously part of the main text, or reside somewhere in between. A prominent example of variation between Bibles are the descriptive titles the Book of Psalms, which in some translations are missing completely, or realized only partially or as part of the title

Century	Language				Overall
	Dutch	English	German	Swedish	
1300–	1361 *Hernse*	1395 *Wycliffe's*			2
1400–	1477 *Delftse*		1460 *Mentelin* SVNT 1478 *Kölner* SVNT		3
1500–	1528 *Vosterman* 1542 *Liesvelt* 1548 *Leuvense* 1560 *Biestkens* 1562 *Deux Aes*	1535 *Coverdale* 1539 *Great* 1560 *Geneva* 1568 *Bishops'*	1534 *Luther*	1541 *Gustav Vasas*	11
1600–	1648 *Lutherse* 1657 *Statenvertaling*	1611 *King James Version*	1662 *Mainzer* SVNT		4
1700–	1796 *Van Hamelsveld* SVNT	1750 *Challoner's*	1781 *Rosalino* SVNT	1703 *Karl XII:s*	4
1800–	1894/1911 *Professoren*	1833 *Webster's* 1890 *Darby*	1871 *Elberfelder*		4
1900–	1939 *Canisius* 1951 *NBG* © 1975 *Willibrord* © 2004 *NBV* ©	(contemp. English) ©	(contemp. German) ©	1917 *års kyrkobibel* 2015 *Folkbibeln* ©	8
Overall	15	10	7	4	36

The '©' symbol marks those translations whose original edition is known (or strongly suspected) to still be under copyright. For the contemporary English and German endpoints, we are still in a dialogue with copyright holders. 'SVNT' marks new electronic editions, created for the purpose of the EDGeS corpus.

Table 1: Selection of Bible translations, per language and century.

(2 a, English, 1535); and which may remain unnumbered as a subtitle before verse one (2 b, German, 1871), or receive verse number one (2 c, Swedish, 1917).

(2) (a) The XLIIII. A psalme of the children of Corah.
MY hert is dyting of a good matter, J speake of that, which J haue made of the kynge: My tonge is the penne of a ready wryter.

(b) Der 45. Psalm.
Dem Vorsänger, auf Schoschanim, für die Söhne Korahs, eine Unterweisung, ein Lied von der Geliebten.
Es wallet mein Herz von gutem Worte. 1
Ich sage: Meine Gedichte dem Könige! meine Zunge sei der Griffel eines fertigen Schreibers!

(c) 45 PSALMEN
För sångmästaren, efter »Liljor»; av Koras 1
söner; en sång, ett kväde om kärlek.
Mitt hjärta flödar över av sköna ord; jag säger: 2
min dikt gäller en konung; en snabb skrivares penna är min tunga.

Notes, cross-references and accompanying markers in the text that indicate their presence are not extracted, and neither are elements like title pages, prologues, registers, etc.[5]The verse numbers themselves are kept to form the basis of the versification information for that particular Bible, but they are not included in the extracted text.

The Bible translations in our corpus, including the oldest ones, come from the start with a division into books and chapters that is largely intercompatible – even though the division is not completely constant: books may be missing, material may be missing from books or moved to separate books, and occasionally the beginning of a chapter occurs in another place in the text. The division of chapters into numbered verses, on the other hand, first appears in the second half of the 16th century for the Bibles in our corpus. Earlier Bible translations have received post-hoc versification. The source of this annotation varies between Bible translations. For instance, our version of *Wycliffe's Bible* (English, 1395) received its verse numbers in the 19th century scientific edition that was the source of our electronic text. The early Dutch Bibles digitized by the SVNT received their verse numbering as part of the digitization. In the case of the *Hernse bijbel* (Dutch, 1361), versification was done in the context of our project. Although we are not dependent on the precise numbers of the verses in a Bible in our corpus, we *do* depend on the division into verses itself, as the corpus is aligned at verse level.

An interesting complication is the presence of divergences between resets in the verse numbering and chapter divisions: in multiple locations and different translations, verse num-

[5]As a rough approximation, one may say that we primarily include everything that is typographically in-line with the body text of a Bible book.

```
Ps.44.0→|The XLIIII. A psalme of the children of Corah.↩
Ps.44.1→|MY hert is dyting of a good matter, J speake of that, which J haue made of the
                                        kynge: My tonge is the penne of a ready wryter.↩

Ps.45.0→|Der 45. Psalm.↩
Ps.45.0→|Dem Vorsänger, auf Schoschanim, für die Söhne Korahs, eine Unterweisung, ein Lied
                                                                    von der Geliebten.↩
Ps.45.1→|Es wallet mein Herz von gutem Worte. Ich sage: Meine Gedichte dem Könige! meine
                                        Zunge sei der Griffel eines fertigen Schreibers!↩

Ps.45.0→|45 PSALMEN↩
Ps.45.1→|För sångmästaren, efter »Liljor»; av Koras söner; en sång, ett kväde om kärlek.↩
Ps.45.2→|Mitt hjärta flödar över av sköna ord; jag säger: min dikt gäller en konung; en
                                                    snabb skrivares penna är min tunga.↩

Eccl.8.17→|...dat hijt wete, hi en salt nochtans niet moghen vinden,↩
Eccl.9.1⟶|Dese dingen heb ic al ouerleit in mijnder herten, om dat ict naerstelijc soude
                                                                        vinden↩
Eccl.9.0⟶|¶ Dat die rechtuaerdighe ende die wijse ende haer wercken in Gods hant zijn.↩
Eccl.9.0⟶|¶ Dat .ix. Capittel.↩
Eccl.9.1⟶|REchtuaerdighe zijnder ende wijse, ende haer wercken zijn in Gods hant
     nochtans so en weet die mensche niet, oft hi des torens, oft der liefden waerdich is↩
Eccl.9.2⟶|Maer alle dingen worden onseker...
```

Figure 1: Opening of Psalm 45 (44 in Vulgate numbering) in three translations, and Ecclesiastes 9:1 interspersed with a chapter heading and summary.

ber 1 precedes or follows a chapter start by several verses. In the following example from the *Vorstermanbijbel* (Dutch, 1528), the chapter break (heading and summary marked with '¶', the chapter start signalled by an initial, reproduced here in bold) even happens *within* a verse.

(3) [17 ...] dat hijt wete, hi en salt nochtans niet moghen vinde*n*, [1] Dese dinge*n* heb ic al ouerleit in mijnder herte*n*, om dat ict naerstelijc soude vinde*n*

¶ Dat die rechtuaerdighe ende die wijse ende haer wercken in Gods hant zijn. ¶ Dat .ix. Capittel.

REchtuaerdighe zijnder en*de* wijse, ende haer wercke*n* zijn in Gods hant nochtans so en weet die mensche niet, oft hi des torens, oft der liefde*n* waerdich is [2] Maer alle dinge*n* worde*n* onseker [...]

In this case, the chapter division is from the printed Bible, but the verse numbering (underlined, within square brackets) was added during creation of the electronic edition, on the basis of the numbering of the Stuttgart Vulgate.

As the examples in this section illustrate, the book-chapter-verse division does not provide us with simple, historically stable 'unique keys'. We therefore use the more flexible mechanism of n–m alignments of bitexts to align our corpus. This is, of course, the common strategy in most work in aligned corpora, but deviates for instance from the approach of Mayer and Cysouw (2014), who place texts into a single, predefined book-chapter-verse grid. More discussion of the alignment method follows below.

Since we have no need for structural markup or layout information, we converted all Bible translations to a simple tab-separated text-based format. As mentioned, we keep the numbering as present in (the digital version of) each translation. Included material that falls outside of a numbered chapter or verse is assigned to a dummy chapter/verse number 0. For our inventory of books, we use the list compiled as part of the OSIS XML scheme.[6]This inventory is fine-grained enough to allow us to differentiate between different source texts for books, in those cases that this leads to considerable differences. For instance, we can distinguish the two versions of the Book of Tobit, which are different enough in length to pose a problem for automatic alignment. Likewise, the books of Esther and Daniel are given different codes depending on whether the additions to these books are integrated (Esther-Greek/Daniel-Greek, following the Greek *Septuagint*), or separated (Esther/Daniel, plus Additions to Esther/Daniel).

After conversion, the examples in 2 and 3 above, look as in Figure 1. Note that we solve the split verse by repeating the verse identifier, which therefore cannot be considered to be unique. Also, the chapter break is independent of the chapter in the verse identifiers, which solves the problem of chapter breaks and verse resets not coinciding.

All texts were aligned at the verse level to the *Nieuwe Bijbel Vertaling* (henceforth NBV, Dutch, 2004), which contains the 66 books of the contemporary protestant canon as well as 11 further apocryphical/deutero-canonical books. The NBV translation is used as a pivot, which simplifies comparing more than two translations of the same verse. Further bitext alignments are calculated from transitive links via the NBV pivot. We used Moore's (2002) method and software[7] to

[6]See `https://wiki.crosswire.org/OSIS_Book_Abbreviations`, consulted Nov 2019

[7]Perl code available from `http://research.microsoft.com` under the name 'Bilingual Sentence Aligner'; consulted: June 2019.

verse-align books, with modifications so as to allow 0–1, 1–1, 1–2 and 2–2 alignments. Manual inspection of a sample of the alignments suggest that this is enough in most cases. As to be expected, the overwhelming majority of alignments are 1–1. The 2–2 alignments are useful for cases where the border between two verses has been put in different places in two translations (Figure 2a). A frequent mistake is the use of spurious 2–2 alignments, where two 1–1 matches would have been correct (Figure 2b). However, for our purposes this is a harmless mistake. The 0–1 and 1–2 (and vice versa) alignments capture cases where a verse is missing from one of the translations, or where two verses are combined into one (Figures 2c and 2d) – these cases *may* be reflected in the NBV numbering, as in the examples. We note that not all verse correspondences are properly captured by these alignment types. For instance, Figure 2e shows how a 2–1 alignment followed by a 1–0 alignment is used for what is best described as a 3–1 verse alignment.

A situation that is not handled very well by Moore's method is when parts of a book are missing from one of the aligned texts. Instead of showing the missing stretch as a contiguous block of 0–1 alignments surrounded by properly aligned material, the aligments start to deteriorate already before and after the missing part. The result is an alignment that is poor overall. A sentence alignment method that explicitly models gaps would probably be a good solution to this problem, but we have not investigated this for the current release of the EDGeS corpus. Fortunately, however, this problem is not very frequent, and books that are problematic in a given language pair are easily recognizable by looking at the length of the books and inspecting verse numbering, so that we can flag bitext alignments that are expected to be problematic.

5. Availability

A recurring problem with corpora compiled from Bible translations is the matter of distribution rights. Many modern translations are still covered by copyright. It is common for the translating bodies to rely on income from selling publishing rights, and these modern translations are therefore rarely found under open licenses. Even available historical translations may not be free to distribute, for instance because the creator of the electronic edition released it under a restrictive license, or – more rarely – because the publication rights are restricted by national regulations (for instance in the case of the royal prerogative with respect to publishing the Authorized/King James Version in the United Kingdom).

Although the lack of redistribution rights need not impede project-internal use, it obviously does limit the value and usefulness of a corpus for the wider research community, which is best served by a corpus that is easily obtainable, and whose license allows researchers to distribute annotations, enhancements and derived works.

License issues have plagued earlier projects to different degrees: Resnik et al. (1999) write they are 'optimistic about [...] making our annotated versions available' (p143), but as a minimal distribution strategy propose to release conversion and annotation scripts. The problematic copyright situation is noted in both Chiarcos et al. (2014) and Mayer and Cysouw (2014). As far as we are aware, the Bible subcorpus of the former has not been made available to the wider research community, whereas the latter has released open subparts as well as derived data for the whole collection. Parts of the parallel Bible corpus used in Kalouli et al. (2018) are released, although under licenses that do not allow further redistribution. Christodouloupoulos and Steedman (2015) do not discuss licensing issues, and the annotated XML files are all downloadable under a CC0 license.[8] The Biblia Medieval project is itself the creator of its electronic editions, and the published Bibles are freely accessible at the project's website (Enrique-Arias and Pueyo Mena, 2008–). For our research purposes, it is important to have the contemporary stages of the languages we investigate, so we have chosen to even include the copyrighted texts from the 20th and 21th century in our corpus. Our take on the problem of redistribution rights has been to create a division in our corpus, with two different modes of availability. The *whole* corpus will be accessible through the OPUS search facilities (Tiedemann, 2012)[9]. For the restricted materials, we have secured permission from, or are in the process of negotiating with, rights holders to make these materials available in the search interface *behind a login*. A strong advantage of the OPUS interface is that is does not only handle bitexts, but allows querying multiple parallel texts simultaneously using the standard CQP query language. We are also confident that the infrastructure around OPUS is sufficiently stable to ensure the materials' accessability for the forseeable future. In addition, the subset of the corpus that we are allowed to redistribute under a permissive license, will be available for download, as will be documentation and conversion code. We refer to the project website `spraakbanken.gu.se/en/projects/complex-verb-constructions` for more information and links to download locations and the material in OPUS.

6. Conclusions

This paper described the development of the EDGeS Diachronic Bible Corpus: a diachronically and synchronically parallel corpus of Bible translations in Dutch, English, German and Swedish. We have explained its design principles, which are driven by the longitudinal and contrastive studies that we wish to perform on the material, and discussed some of the challenges presented by the variation present in a collection of Bible translations from such a long time period. The corpus will be available in two forms: the whole corpus will be accessible for researchers behind a login in the OPUS search infrastructure, the open parts of the corpus will be available for download. After the release of the corpus, future work will concentrate on adding linguistic annotation to support linguistically motivated queries.

[8] See `github.com/christos-c/Bible-corpus`; consulted 14 Nov 2019.

[9] `http://opus.nlpl.eu/`

|Deut 1:32| Desondanks vertrouwde u niet op de HEER, uw God, |1:33| hoewel hij u voorging op uw weg || om een plaats voor u te zoeken waar u uw kamp kon opslaan, en u 's nachts met een vuur en overdag met een wolk de weg wees die u moest gaan.

|1:32| ende noch aldus so en hebdi niet gheloeft den heer uwen god || die voer v ghegaen is inden wech |1:33| ende die stat wijsde daer gi uwe tenten slaen sout: ende die v des nachtes den wech wijsede ende thoende bij vyer: ende des daghes bider columnmen eenre wolken.

(a) correct 2–2 alignment; '||' mark hypothetical, projected boundaries (2004 NBV–1447 Delftse)

|Prov 20:5| Wat omgaat in een mensenhart is als diep verborgen water, iemand met inzicht brengt het naar boven. |20:6| Velen roemen hun eigen trouw, maar wie vindt een mens die werkelijk betrouwbaar is?

|20:5| Ghelijc diepe wateren also is die raet in eens mans herte: mer die wijse man sal dien wtsceppen. |20:6| Veel menschen heetmen ontfermhertich mer wie sal vinden enen ghetrouwen man?

(b) spurious 2–2 alignment; two 1–1 beads would have been correct (2004 NBV–1447 Delftse)

|Rom 9:11-12| en al voor ze geboren waren en nog niets goeds of slechts hadden gedaan, werd haar gezegd: 'De oudste zal de jongste dienen.' Gods besluit blijft namelijk van kracht: God kiest een mens niet uit op grond van zijn daden, maar omdat hij hem roept.

|9:11| selbst als die Kinder noch nicht geboren waren und weder Gutes noch Böses gethan hatten, (auf daß der Vorsatz Gottes nach Auswahl bestände, nicht aus Werken, sondern aus dem Berufenden) |9:12| ward zu ihr gesagt: "Der Größere wird dem Geringeren dienen";

(c) correct 2–1 alignment; reflected in NBV verse numbers (2004 NBV–1871 Elberfelder)

|John 5:3| Daar lag een groot aantal zieken, blinden, kreupelen en misvormden.

|5:3| In these lay a great multitude of sick, of blind, of lame, of withered: waiting for the moving of the water.

|5:4| And an angel of the Lord descended at certain times into the pond and the water was moved. And he that went down first into the pond after the motion of the water was made whole of whatsoever infirmity he lay under.

|5:5| Er was ook iemand bij die al achtendertig jaar ziek was.

|5:5| And there was a certain man there that had been eight and thirty years under his infirmity.

(d) correct 0–1 alignment in context; reflected in NBV verse numbers (2004 NBV–1780 Challoner's)

|Exod 40:12| Laat dan Aäron en zijn zonen naar de ingang van de ontmoetingstent komen en reinig hen met water.

|40:12| And thou shalt bring Aaron and his sons to the door of the tabernacle of the testimony, and having washed them with water,

|40:13| Trek Aäron de heilige kleding aan en zalf hem; zo heilig je hem om mij als priester te dienen. |40:14| Ontbied zijn zonen, trek hun de tunieken aan

|40:13| Thou shalt put on them the holy vestments, that they may minister to me, and that the unction of them may prosper to an everlasting priesthood.

|40:15| en zalf hen zoals je hun vader gezalfd hebt; dan kunnen ook zij mij als priester dienen. Door deze zalving wordt hun voor altijd, voor alle komende generaties, het priesterschap verleend.'

|40:16| Mozes deed alles wat de HEER hem had opgedragen.

|40:14| And Moses did all that the Lord had commanded.

(e) actual 3–1 correspondence in context, aligned as a sequence of 2–1, 1–0 (2004 NBV–1780 Challoner's)

Figure 2: Examples of different alignment types

7. Acknowledgements

The corpus presented in this paper is constructed in the context of the project 'The rise of complex verb constructions in Germanic' (Swedish Research Council grant no. 2017-01848, PI: Evie Coussé). The design and the first compilation phase of the EDGeS corpus was carried out by Dirk-Jan de Kooter and Trude Dijkstra under supervision of Nicoline van der Sijs at the Meertens Institute, Amsterdam.

8. Bibliographical References

Beelen, H. and van der Sijs, N. (2014). Crowdsourcing de bijbel. *Neerlandia*, 118(2).

Breder Birkenes, M., Fleischer, J., and Leser-Cronau, S. (in press). A diachronic and areal typology of agreement in Germanic. *STUF – Sprachtypologie und Universalienforschung*. To appear 2020.

Chiarcos, C., Sukhareva, M., Mittmann, R., Price, T., Detmold, G., and Chobotsky, J. (2014). New technologies for Old Germanic. Resources and research on parallel Bibles in older Continental Western Germanic. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 22–31, Gothenburg, Sweden, April. Association for Computational Linguistics. doi:10.3115/v1/W14-0604.

Christodouloupoulos, C. and Steedman, M. (2015). A massively parallel corpus: the Bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395. doi:10.1007/s10579-014-9287-y.

Coussé, E. (2015). Constructional complexification. The rise of double modal constructions in Dutch. *Taal en Tongval*, 67:149–167.

Cysouw, M. and Wälchli, B. (2007). Parallel texts: using translational equivalents in linguistic typology. *STUF – Sprachtypologie und Universalienforschung*, 60(2):95–99. doi:10.1524/stuf.2007.60.2.95.

Dagson, J. (2013). *Vilken bibel*. XP Media, Handen, Sweden.

de Vries, L. (2007). Some remarks on the use of Bible translations as parallel texts in linguistic research. *STUF – Sprachtypologie und Universalienforschung*, 60(2):148–157. doi:10.1524/stuf.2007.60.2.148.

Enrique-Arias, A. and Pueyo Mena, F. J. (2008–). Biblia Medieval. `http://www.bibliamedieval.es`.

Haug, D. T. T. and Jøhndal, M. L. (2008). Creating a parallel treebank of the Old Indo-European Bible translations. In C. Sporleder et al., editors, *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 27–34, Marrakesh. `http://www.lrec-conf.org/proceedings/lrec2008/workshops/W22_Proceedings.pdf#page=31`.

Hilpert, M. (2011). Grammaticalization in Germanic languages. In Bernd Heine et al., editors, *The Oxford Handbook of Grammaticalization*. Oxford University Press, Oxford. doi:10.1093/oxfordhb/9780199586783.001.0001.

Kaiser, G. A. (2005). Bibelübersetzungen als Grundlage für empirische Sprachwandeluntersuchungen. In C. D. Pusch, et al., editors, *Romanische Korpuslinguistik II: Korpora und diachrone Sprachwissenschaft*, number 130 in ScriptOralia, pages 71–83. Gunther Narr Verlag, Tübingen.

Kalouli, A.-L., Kaiser, K., Hautli-Janisz, A., Kaiser, G. A., and Butt, M. (2018). A multilingual approach to question classification. In N. Calzolari, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2715–2720, Miyazaki, Japan. European Language Resources Association (ELRA). `http://www.lrec-conf.org/proceedings/lrec2018/pdf/13.pdf`.

Mayer, T. and Cysouw, M. (2014). Creating a massively parallel Bible corpus. In N. Calzolari, et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik. European Language Resources Association (ELRA). `http://www.lrec-conf.org/proceedings/lrec2014/pdf/220_Paper.pdf`.

Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In S. Richardson, editor, *Machine Translation: From Research to Real Users*, pages 135–144, Berlin, Heidelberg. Springer. doi:10.1007/3-540-45820-4_14.

Pettersson, J. (2017). Nordic Bible translations in medieval and early modern Europe. In W. François et al., editors, *Vernacular Bible and Religious Reform in the Middle Ages and Early Modern Era*, pages 107–150. Peeters, Leuven.

Resnik, P., Olsen, M. B., and Diab, M. (1999). The Bible as a parallel corpus: Annotating the 'Book of 2000 tongues'. *Computers and the Humanities*, 33(1):129–153. doi:10.1023/A:1001798929185.

Rosemeyer, M. and Enrique-Arias, A. (2016). A match made in heaven: Using parallel corpora and multinomial logistic regression to analyze the expression of possession in Old Spanish. *Language Variation and Change*, 28(3):307–334. doi:10.1017/S0954394516000120.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In N. Calzolari, et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA). `http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf`.

Wansbrough, H. (2008). History and impact of English Bible translations. In M. Sæbø, editor, *Hebrew Bible / Old Testament: The history of its interpretation*, volume II: From the Renaissance to the Enlightenment. Vandenhoeck & Ruprecht, Göttingen.