Cambridge Semitic Languages and Cultures

# Linguistic Theory and the Biblical Text

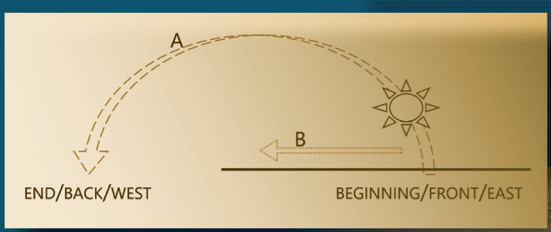
Edited by William A. Ross and Elizabeth Robar

Cognitive Linguistic Theory

Functional Grammar

οὗτος ἦν ἐν ἀρχῇ πρὸς τὸν θεόν. πάντα δι' αὐτοῦ ἐγένετο, καὶ χωρὶς αὐτοῦ ἐγένετο οὐδὲ ἕν ὃ γέγονεν. ἐν αὐτῷ ζωὴ ἦν, καὶ ἡ ζωὴ ἦν τὸ φῶς τῶν ἀνθρώπων· καὶ τὸ φῶς ἐν τῇ σκοτίᾳ φαίνει, καὶ ἡ σκοτία αὐτὸ οὐ κατέλαβεν.

Historical Linguistics




Complexity Theory

Generative Linguistics

This was in the beginning with God. All things through him came to be, and outside of him came to be nothing that came to be. In him was life and the life was the light of people, and the light shines in the darkness and the darkness did not comprehend it (Jn. 1:2–5)

Pragmatics of Information Structure

Computational Linguistic Analysis

UNIVERSITY OF CAMBRIDGE
Faculty of Asian and Middle Eastern Studies

https://www.openbookpublishers.com





William A. Ross and Elizabeth Robar (eds), *Linguistic Theory and the Biblical Text*. Cambridge, UK: Open Book Publishers, 2023, https://doi.org/10.11647/OBP.0358



Semitic Languages and Cultures 20.

ISSN (print): 2632-6906
ISSN (digital): 2632-6914
ISBN Paperback: 978-1-80511-108-5
ISBN Hardback: 978-1-80511-109-2
ISBN Digital (PDF): 978-1-80511-110-8
DOI: 10.11647/OBP.0358

Cover image: A section of Cisneros' original complutensian polyglot Bible, https://en.wikipedia.org/wiki/File:Cisneros%27_original_complutensian_polyglot_Bible_-2.jpg; additional text and diagrams created by authors.

Cover design by Jeevanjot Kaur Nagpal

The main fonts used in this volume are SIL Charis, SBL Hebrew, and SBL Greek.

# COMPUTATIONAL LINGUISTIC ANALYSIS OF THE BIBLICAL TEXT

*Willem Th. van Peursen*

## 1.0. History and Development of the Theory

### 1.1. Introduction

The application of computational linguistics to the Bible is part of the broader field of 'Bible and Computer' as it was coined in the 1970s and which encompasses, besides linguistic research, an ever-increasing field of computational textual analysis applied to the biblical text. It takes place at the intersection of biblical studies and the rapidly developing field of Digital Humanities.

This chapter deals with computational linguistics as a method, besides the other methods described in this volume. It should be recalled, however, that computational linguistics is interwoven with other approaches. When we compare the various available syntactic databases of the Hebrew Bible, we can observe, for example, that of the three most well-known databases the Andersen–Forbes database is explicitly eclectic in its linguistic theory (cf. Andersen and Forbes 2012).[1] On the other

[1] For project documentation and bibliographical references, see http://andersen-forbes.org, accessed 1 May 2023.

 https://doi.org/10.11647/OBP.0358.05

hand, the ETCBC database or BHSA (cf. Kingham and Van Peursen 2018) builds on the form-to-function approach developed by Jaap Hoftijzer and Wolfgang Richter (Van Peursen 2007, 140–41) and is influenced by the text-syntactic approach developed by Harald Weinrich and Wolfgang Schneider (Van Peursen 2020a, 140–55). Holmstedt's and Abbegg's Accordance Hebrew Syntax Database (Holmstedt and Cook 2018) is highly informed by generative linguistics (Accordance documentation 2014).[2]

Because of these observations, it would be an oversimplification to treat computational linguistics as a method distinct from, for example, generative linguistics or Cognitive Linguistics. Each computational linguistic analysis uses a digital corpus and each of these corpora is rooted in linguistic theories (see below, §2.3). Moreover, computational linguistics is a broad field that includes various approaches such as rule-based computer-science, statistics, Artificial Intelligence, and Deep Learning.[3] All these approaches have been applied to the biblical languages, and hence this chapter will present various approaches rather than one single method. Nevertheless, we shall see that these computational approaches have some common features that justify treating them together and that they have developed further in ways that are typical for computational corpus linguistics and go

[2] For other database projects on Biblical Hebrew that were active over the last decades see Kroeze (2013).

[3] In addition, computational linguistics as a discipline also covers approaches such as speech recognition or natural language generation that fall outside the scope of the current chapter. For an overview see, e.g., Clark, Fox, and Lappin (2010); Jurafsky and Martin (2021).

beyond the linguistic theories underlying the annotations in the respective databases. This reality justifies a separate chapter in the current volume devoted to computational linguistics as a method by itself.

## 1.2. The Beginnings

From our remarks in the introduction, it will be evident that it is impossible to give a historical survey of the application of computational linguistics to Biblical Hebrew without considering the wider context. This is the context of text and computing as an emerging field of studies in the twentieth century. Leaving aside for the moment predecessors such as the mechanical machines that were made, or at least designed, in the nineteenth century, such as the design for a mechanical general-purpose computer by Charles Babbage, we will start this survey with the emergence of the forerunners of the modern computers in the 1940s and 1950s. In this period, we see the transformation of the calculation machine into the universal machine: that is, a machine that can do any task for which it is programmed. In these early years, Robert Busa started his famous project of the *Index Thomisticus*, which involved the complete lemmatisation of the works of Thomas Aquinas (which consists of 181 works, comprising 11 million words). This monumental project started in 1949 and lasted about thirty years.

The first universal computers created were not primarily meant for text processing. It should be recalled that the combination of text/language and computation/calculation is not as self-evident as it now seems. Even long before the emergence of

computers, the fuzziness and ambiguity of natural language frustrated projects like those of Wilhelm Leibniz (1646–1716), who for "his whole life… continued to believe in the construction of a language consisting of logical symbols that could be manipulated by means of a calculator. Such a language, and a machine to 'calculate' it, would enable any philosophical debate to be settled with the click of a button" (Van der Weel 2011, 106). Likewise, "around a hundred years ago, polymaths like Bertrand Russell were furiously fighting to capture the nuances of language with a view to developing a universal formal language," which remained an ongoing academic pursuit that continued in the field of computer science, but appeared to be a highly challenging project (see action.ai 2021).

It was only in the 1960s and the 1970s that the marriage between computer and text took place. In the 1960s, computers became able to process text. A milestone was the first edition of the ASCII standard in 1963. This standard involved a 7-bit encoding in which, for example, 1000000 stands for @, 1000001 for A, and 1000010 for B, and 1000011 for C. In total, the ASCII standard contained 128 codes. Accordingly, the first attempts to create electronic versions of the Hebrew Bible had to accommodate this standard.

These attempts started in the 1970s. In 1970, Francis Andersen and Dean Forbes started a project that finally resulted in the Andersen–Forbes database. In the same year, Christof Felix Hardmeier (1970) from Greifswald reported on his own experiments in his article on the new potential of electronic data processing. Somewhat later, in 1977, the Werkgroep Informatica

Vrije Universiteit (WIVU) was established in Amsterdam under the guidance of Eep Talstra (after whom the WIVU was rebaptised as the Eep Talstra Centre for Bible and Computer [ETCBC] in 2013), which marked the start of the WIVU/ETCBC database. At Westminster Theological Seminary in Philadelphia, J. Alan Groves started pioneering work which initiated the research at what is now called the J. Alan Groves Center for Advanced Biblical Research. This work resulted in the Westminster Leningrad Codex (first released in 1987), to be followed by the Westminster Hebrew Morphology (also known as the Groves-Wheeler morphology) and the Westminster Hebrew Syntax.

Pioneers such as Andersen, Forbes, Hardmeier, and Talstra found each other in the *Association Internationale Bible et Informatique* (AIBI), which was established in 1982 and held its first conference in 1985 in Louvain-la-Neuve. Besides the pioneers already mentioned (and others such as Marc Vervenne or Emanuel Tov), a driving force behind this organisation was R. F. Poswick from the Benedictine monastery of Maredsous. The theme of the first AIBI conference was 'the text', and that was precisely the main challenge during those years: how to represent the Hebrew text and linguistic annotations. There was no Unicode, no markup language like HTML and XML, and not even a PC back then. The first challenge these pioneers faced was building electronic text corpora, displaying them on the screen, and handling the right-to-left writing direction for Hebrew.

## 1.3. Interface, Office and Network

Major changes took place in the 1980s and 1990s, which were related to such terms and abbreviations as GUI (Graphical User Interface), the DTP (Desktop Publishing) revolution, and WYSISWYG (What You See Is What You Get). These changes can be illustrated by the introduction of the Apple Lisa in 1983, the first version of the program PageMaker in 1984, and the first release of Microsoft Office in 1989. These developments marked a change in the application of the computer towards more office-related activities. With this development, the use of the computer became much more widespread, both in number of users and in types of applications. In the field of biblical studies this resulted in the appearance of software packages such as BibleWorks (first release in 1992) and Accordance (first release in 1994).

These new tools became extremely helpful for biblical scholars. One could now display the Hebrew Bible and the ancient versions side by side, search for words and word combinations in the electronic text instead of consulting a printed concordance, and store large commentaries on one's disk rather than on one's bookshelves.[4] A side-effect of this development, however, was a shift of focus. The early pioneers of 'Bible and Computer' were mainly concerned with the computer as an analytical tool, but in practice, it rather became a useful office tool. Being able to search for a word with a query instead of looking

[4] But often, again, the computer was used to generate concordances that were published in print. Thus, e.g., Postma, Talstra, and Vervenne (1983); cf. Oosting (2016, 195).

it up in a printed concordance may be a little bit faster, but it is not a methodological improvement. A burning question that occupied the early pioneers but seemed to be hardly a concern for the broader community of biblical scholars was: How can we go beyond the imitation of the traditional instruments?[5]

Another effect of the developments in the 1980s described here was that some of the databases that were initiated in the 1970s and 1980s became commercial products. To my best knowledge, it is only the ETCBC database that is publicly available,[6] while the Andersen–Forbes database is only available in the commercial Bible software packages of Logos and Accordance, and the more recent Holmstedt–Abbegg database only in Accordance.[7] This has hindered further development within the scholarly community, because one of the primary conditions of

---

[5] See the telling title of Talstra and Dyk (2006): 'The Computer and Biblical Research. Are there Perspectives beyond the Imitation of Classical Instruments?'

[6] https://github.com/ETCBC/bhsa, accessed 4 May 2023. Recently, also the MACULA Hebrew syntax trees have become available at https://github.com/Clear-Bible/macula-hebrew, accessed 4 May 2023. These syntax trees have been developed by Clear Bible, Inc. together with the Groves Cente and build on the Westminster Hebrew Syntax Without Morphology and the Open Scriptures Hebrew Bible morphology (serving in place of the Westminster Hebrew Morphology). The Groves Center has also released the Westminster Hebrew Syntax Without Morphology at https://github.com/Clear-Bible/macula-hebrew/tree/main/sources/GrovesCenter, accessed 4 May 2023.

[7] Cf. Accordance documentation (2014), for the advantages that Holmstedt and Abbegg considered for integrating their database into the Accordance software right from the start.

computational linguistics is that the analyses are retrievable and that the underlying algorithms are available on online platforms such as GitHub (cf. below, §4.3).

In the 1990s, a new element radically changed the digital landscape: networking. The World Wide Web was launched in 1991 and in the same year the first version of Unicode was released. These two milestones were closely related, because only with the unequivocal definition of characters in Unicode was it possible to exchange text that remained stable regardless of the environment in which it was read. For PC users, the internet became accessible through the browsers that came onto the market, such as Netscape Navigator in 1994 and Internet Explorer in 1995.

This new development was also soon picked up by biblical scholars. Electronic journals in the field of biblical studies were initiated, such as *TC: A Journal of Biblical Textual Criticism* and the *Journal of Hebrew Scriptures*, which both started in 1996.[8]

At the turn of the century, a new stage started with the introduction of more interactive forms of publication and communication, in which the users became both consumers and contributors and in which the dividing line between information consumption and information creation was blurred. This is often labelled 'Web 2.0'. Milestones include the launch of Wikipedia (2001) and the emergence of social media such as Facebook (2004) and Twitter (2006).

---

[8] Mention should also be made here of *Hugoye*, a journal in the field of Syriac studies, which started in 1998.

This ever-growing field of textual and social computer applications affected biblical studies. The use of electronic tools was no longer the privilege of biblical scholars. More and more, everyone had an increasing number of online Bibles and Bible study tools at their disposal. Likewise, the field of 'Bible and Computer', as defined by the AIBI, was expanding as well. At the sixth AIBI conference held in Stellenbosch in 2000, there were sections on grammar, statistics, and discourse, but also on education, multi-media, publishing, and community, all in relation to the Bible and the computer.

## 1.4. Reorientation: Methodological Innovation?

The development described above was not the programme that the pioneers of the 1970s and 1980s had in mind when they started their work. A re-orientation took place in the first decade of the twenty-first century. The seventh AIBI conference (convened by Marc Vervenne, Leuven, 2004) and the eighth conference (convened by Luis Vegas Montaner, Madrid, 2008) were both presented as expert meetings focusing on the question of how the computer can play an innovative role in biblical scholarship.[9] How could the computer be used as an analytical tool, rather than merely as a library, an office tool, and an imitation of traditional tools, which it apparently had become in the 1990s?

The question regarding the role of the computer in methodological innovation touched upon the more encompassing

[9] See the overview given in Poswick (2010), but note that the Leuven 2004 conference is absent from Poswick's overview.

question regarding textual scholarship as a humanities discipline in relation to computer science, which typically belongs to the sciences. Computation and the related scientific mode of inquiry gave the ability to sort, quantify, reproduce, and report text, but how could this be fruitfully combined with interpretation as the valued mode of assigning or discovering meaning as understood in traditional scholarship and the related reflexive concepts of individualism and subjectivity (Van Peursen 2010)?

The final decades of the twentieth century had witnessed a shift in the humanities from the hermeneutic and critical tradition of the nineteenth and twentieth centuries towards the identification and representation of patterns by digital means in the second half of the twentieth and the early years of the twenty-first century. Rens Bod (2013) coined the two phases 'Humanities 1.0' and 'Humanities 2.0'. (Note that '2.0' is used here differently to in 'Web 2.0' discussed above). Humanities 1.0 embodied the traditional understanding of the humanities as it was framed at the end of the nineteenth century. Wilhelm Dilthey (1833–1911) and others advocated a clear-cut distinction between the humanities and the sciences, the first mainly involved in *Verstehen* (understanding) the second in *Erklären* (explanation). This distinction had a significant impact on modes of scholarship, but also on the organisation of academia, where most institutions have separate departments for the humanities and the sciences. With the appearance of the computer as a tool for textual scholarship (Humanities 2.0), this distinction was blurred. How could this distinction be maintained now that computer scientists seemed to be analysing texts in the same way in which natural

scientists analysed DNA structures? Rens Bod (2013, 177) has argued that this new mode of scholarship should not be the end point, but that a next step should be taken (which he labelled Humanities 3.0), in which Humanities 1.0 and Humanities 2.0 are combined and in which the hermeneutic and critical tradition of Humanities 1.0 should be applied to the tools and patterns obtained by Humanities 2.0.

## 1.5. From Rule-Based Analysis to Machine Learning

While biblical scholars and textual scholars in general were busy incorporating computer science into their disciplines, computer science itself developed further with astonishing speed. Let us illustrate this with the example of machine translation. From the early days of computational linguistics, it was evident that it would be tremendously useful if the computer could be used to translate a text from one language into another. As early as the 1970s, attempts were made to achieve this task by rule-based machine translation. In this approach, the input that the computer receives is the text to be translated and language rules. These rules include, for example, a bilingual lexicon, morphology, and syntax. The more refined those rules, the fewer errors the translation contains and the better it becomes. However, after decades of improvements, the results did not meet the high expectations. Natural language appeared more unruly than people thought (cf. above, §1.1). The rule-based techniques of the 1970s to the 1990s were replaced by statistical approaches in the 1990s until the 2010s. However, although there was significant progress, the real breakthrough came only with the

application of machine learning. Here the input is no longer a text to be translated and a set of rules to carry out this task, but rather a large collection of training data, in this case of parallel texts in two languages, from which the computer itself can learn how to translate. Although the mechanisms that are at work are largely hidden, the performance is outstanding.

If we define machine learning more precisely, we can say that it is the ability to learn without being explicitly programmed. It is a subgroup of Artificial Intelligence, which refers to any technique that enables computers to mimic human behaviour or, more precisely, the effort to automate intellectual tasks normally performed by humans. Artificial Intelligence (AI) went through various stages by itself, from symbolic AI, which was prevalent until the 1980s and involved the application of explicit rules and the manipulation of logic, to machine learning, where the computer goes beyond the instruction and rules it is given and learns by itself how to perform a certain task. A subgroup of machine learning is deep learning, which refers to the extraction of patterns from data with the help of neural networks. In the case of machine learning, we can distinguish between supervised machine learning, in which the machine learns from human-labelled examples, and unsupervised learning, in which the machine has to detect patterns in the unlabelled data by itself. Supervised methods include attempts for text classification. In biblical studies, an example is Dicta (see n. 10), which provides an exciting collection of tools for author recognition and text classification, such as the Tiberias Stylistic Classifier (cf. below,

§§2.5 and 3.3).[10] These tools can be used, for example, to classify a text of debated origin along the lines of early and late Biblical Hebrew.

Experts may challenge the rather simple definitions given here for machine learning, deep learning, and Artificial Intelligence, and there is much debate about the exact nature and definitions of the various designations given here. But the main point to be made is that a major shift has taken place, which affects our whole understanding of using the computer in linguistic analysis. Elsewhere I have suggested that the transformations that are taking place now with the transition from rule-based approaches to machine learning may even mark a more drastic discontinuity with existing methods of biblical interpretation than the appearance of the computer as an exegetical tool in the last decades of the twentieth century (Van Peursen 2020b, 310). Whereas many of the rule-based approaches could somehow mimic traditional approaches (e.g., queries replacing concordances or manually created lists), machine learning opens up completely new avenues of scholarship that may lead to new forms of human-computer interaction in the interpretation of texts.

## 1.6. Corpora and Fuzzy Data

The application of computational linguistics to the Bible implied that the Bible was considered a corpus, and thus it entered the field of corpus linguistics. However, in this field of studies, the

[10] See https://tiberias.dicta.org.il/, accessed 4 May 2023.

Hebrew Bible did not match the corpora from other languages and periods. There are huge differences between, for example, the British National Corpus (BNC) and the Hebrew Bible. The BNC has more than 100 million words, as against 420,000 words in the Hebrew Bible. The BNC has extensive metadata about, for example, author and date of origin, whereas for almost every part of the Hebrew Bible authors and provenance are debated issues. The BNC has been carefully selected to create a representative linguistic corpus, whereas the Hebrew Bible, whatever the selection processes that made it the biblical canon that we now have, was never intended to be linguistically representative and was selected along completely different criteria. Accordingly, at least until the turn of the century, the computational analysis of the Hebrew Bible was a questionable undertaking according to the developing standards of computational linguists. When in the early 1990s Eep Talstra once presented his research on Deuteronomy to an audience of computational linguists, he met with much misunderstanding. How could he study a corpus of which he did not know the date of origin? Unaware of the complex questions regarding sources and editorial processes in the Hebrew Bible that have puzzled biblical scholars for centuries, one of the respondents suggested that Talstra should first clean up his data (that is: stripping it of any later additions so that what remains is a corpus of which the date and provenance are clear) before any linguistic or textual analysis could start (Talstra 2010, 54).

This situation has completely changed since the above-mentioned emergence of Web 2.0. Currently, much linguistic research is conducted on tweets, blogposts, and other digitally

born texts that represent a kind of fuzzy data, with little context and little metadata, all of which resembles the Hebrew Bible much more than the BNC does. Thanks to these developments, computational research into the Bible has found a better connection with the wider field of Digital Humanities than in the last three to four decades of the twentieth century.

## 2.0. Key Theoretical Commitments and Major Concepts

### 2.1. Solid Criteria instead of Subjective Intuition

Biblical scholars were among the first who experimented using the computer as a tool for the study of texts and languages. In the 1970s, at the dawn of computer-aided textual analysis and more specifically in the newly emerging field of 'Bible and Computer', the mission of the pioneers (above, §1.2.) was clear: Make meaningful, substantiated statements about the Bible. Such an effort did not guarantee the correct interpretation, but at least it could identify interpretations that did not match the facts. Traceability and transparency played an important role in this development.[11]

[11] A typical example is Hardmeier's above-mentioned article. Addressing the question as to whether the computer can help in traditional source criticsm, Hardmeier (1970, 180) argues: "Die maschinelle Konkordanzarbeit ermöglicht dagegen ein Zweifaches: Einmal kann der Kriterienkatalog über die Wortschatzstatistik hinaus auf neue, rein formale Struktureigentümlichkeigen bestimmter Texte ausgedehnt werden (…) Zum anderen kan überprüft werden, wieweit lexikalische und formale Merkmale für bestimmte Textschichten überhaupt charakteristisch sind." It was, however, only after the emergence of machine

This mission fit in a positivist, modernist climate, and was especially opposed to an unbridled theologising based on individual words and etymologies, which has made, among others, Kittel's *Theologisches Wörterbuch zum Neuen Testament* (1942–1979) famous (or infamous). In the 1960s, James Barr assessed this approach critically on the basis of general linguistic and philological insights (see Barr 1961). Talstra, who earlier in the 1970s had studied with Barr in Manchester, would certainly agree with Barr's criticism, and started creating a database that was not focused on etymology and semantics, but on syntax.

This mission of those pioneers in Digital Humanities (at that time other labels were used such as Alpha-Informatics) still plays an important role and is reflected in open science practices. Statements about occurrences of words or patterns can be made traceable and reproducible, for example by publishing queries on the SHEBANQ website of the ETCBC.[12] At the same time, we have also seen developments in computational textual analysis that run counter to the ideals of the pioneers (above, §1.3).

## 2.2. Deep Blue and AlphaGo

The introduction of the computer in the workplace of the exegete in the last decades of the twentieth century enabled the biblical scholar to be more systematic, objective, and quantitative. The qualification 'objective' does not mean that computational data

learning that the computer was used fruitfully to address the traditional source-critical questions; cf. below, §3.3.

[12] See https://shebanq.ancient-data.org, accessed 4 May 2023. For the underlying ideas and the role of annotation, see Roorda (2018).

is theory-neutral or that the computer provides the final answer (as it is sometimes misunderstood), but it does indicate another mode of scholarship. The added value of this work lies in its systematic approach, which reduces the role of intuitive *ad hoc* interpretations, on the one hand, and in the increasing complexity of searches and analyses, on the other.

In the 1970s–2000s, the claim that the computer was more systematic and objective was frequently met with scepticism by traditional scholars, who often found those computer guys weird and imagined that using the computer as a tool in biblical studies was in fact building an echoing well: what you get out of the computer depends on what you put into it. Often the question was raised: What does the computer deliver that could not be delivered without it? To parry this criticism, I often used the analogy of the chess computer, which seemed especially apt since Deep Blue had beaten Gary Kasparov in 1997. The mind of the human chess player works efficiently because it recognises patterns and therefore has a useful selection mechanism, whereas the computer, so to speak, calculates everything (still an interesting study is De Groot 1946). However, the speed with which the computer does so (which has increased over the years) is so immense that it surpasses human capacities. Likewise, once you have an annotated database, questions that would take months or years when addressed manually—e.g., a statistical overview of *plene* or defective spellings, or a collection of all clause patterns in the Hebrew Bible where the object precedes the verb (for examples, see §3 below)—can now be answered in very short time periods.

Whereas the analogy of the chess computer worked well until the first years of the twenty-first century, it does not capture the developments that have taken place over the last two decades, described above (§1.5), such as the emergence of machine learning. For these developments, we can rather use the analogy of the Go computer (see Dorobantu 2022). Whereas the human world champion of chess was defeated by the computer in 1997, it took until 2016 before a computer program defeated the human world champion in Go. The difference between chess and Go is that with Go the complexity of the game and the number of possible continuations is exponentially higher than with chess. Moreover, in Go there is also a strong aesthetic aspect. Because of the infinite number of possibilities and the aesthetic aspects, computational calculation power is not enough to win the game. Rule-based approaches (enabled by calculation power) were insufficient, but learning capabilities (enabled by pattern recognition) succeeded. Likewise, in textual analysis the computer is no longer merely a powerful calculation or sorting machine. It has become a much more complex instrument, and to some extent less dependent on human input (above, §1.5).

## 2.3. The Role of Linguistic Theory in the Creation of Text Databases

The different endeavours to create linguistic databases of the Hebrew Bible reflected the different approaches that were and are current in biblical linguistics. Each approach has its advantages and disadvantages. And even in those cases where the builders of a database try to be as theory-neutral as possible, it

will be evident that any database and any choice that is made is informed by one's position about Biblical Hebrew and about language in general. This is not only a challenge for computational approaches. It is the case in any linguistic or textual study of the Hebrew Bible or Greek New Testament.

Let us have a look how linguistic theory functions in the three most well-known databases (above, §1.1; see also Miller-Naudé and Naudé 2018, 7), which we will discuss in the order of their age. First, the Andersen–Forbes database is eclectic and hence somewhat ambiguous in its relation to generative grammar. Andersen and Forbes explicitly reject Chomskyan linguistics but also "find much of value in the work of the generativists, especially generalized phrase structure grammar" (Andersen and Forbes 2012, 14; see also Van Peursen 2015, 301). One of the main reasons for their rejection of Chomskyan linguistics is their claim that Biblical Hebrew belongs to the non-configurational languages, which are "a serious impediment to the transformationalists' quest for Universal Grammar" (Andersen and Forbes 2012, 87).[13] Andersen was also influenced by structuralism and by Kenneth Pike's tagmemics (Miller-Naudé and Naudé 2018, 7). Informed by syntax, function, and semantics, they developed a rich set of annotation labels, including, among others, seventy-six part-of-speech labels.

Second, the ETCBC tried to be more independent of linguistic theory by following the principles of distributional analysis,

[13] For a different view on the question as to whether Biblical Hebrew is a configurational language, partly in response to Andersen and Forbes, see Kaajan (2019).

form-to-function, and bottom-up. Unlike the Andersen–Forbes database, the ETCBC database has a rather minimal parts-of-speech set (Kingham 2018). And unlike the other databases, the ETCBC database does not immediately assign functions to forms, but starts with a distributional analysis of linguistic phenomena (at all levels, for example varying from morphemes to clause patterns) before functional labels are assigned. Moreover, the deduction of functions from formal criteria is transparently and traceably documented in auxiliary files that are part of the data creation workflow (Kingham 2018; Kingham and Van Peursen 2018).

The 'bottom-up' description is used for approaches that start from the identification of morphemes and word level analysis and move from there to the higher levels of phrases, clauses, sentences, and text-syntactic relations. It is often combined with the form-to-function principle, which holds that we should first make a distributional analysis of forms and patterns before any function can be deduced. It starts from the awareness that we know little about the biblical languages and that to avoid creating an echoing well out of our own analysis or database, we should start with observable textual phenomena before we proceed to function or even semantics.

'Bottom-up' is often contrasted with 'top-down'. In Biblical Hebrew linguistics, the latter is represented, for example, in the textlinguistic approach of Robert Longacre (1989), which is much more informed by cross-linguistic evidence and applies categories known from other languages (such as narrative, predictive, hortatory, or expository genres; techniques for, for

example, distinguishing between mainline and offline information or for indicating the peak of a text or discourse). Likewise, the distributional analysis of forms and patterns, which to some extent is like Construction Grammar or exemplar-based syntax as it developed in the 1980s and 1990s, is often considered as a counter-reaction to the generative linguistic framework.

Third, the Holmstedt–Abbegg database, also called the Accordance Hebrew Syntactic Database, is based on a generative framework (cf. Accordance documentation 2014). Whereas the Andersen–Forbes and ETCBC databases started in the 1970s, the Holmstedt–Abbegg database started more recently, in 2008. Their intention was to create a database upon a model of Hebrew syntax that differed from the two existing databases, with "a tight focus on syntax, grounded in (but not bound by) Chomskyan generative linguistic theory" (Holmstedt and Cook 2018, 2).[14] More specifically, they adhered to Chomskyan minimalism, which was developed in the 1990s from the Government-and-Binding model that was prevalent in the 1980s, but they also realised that "to base the database and its underlying tagging scheme on a fully articulated minimalist framework would be inappropriate." For this reason, they combined their adherence to Chomskyan theory with the motto "data primary, theory wise" (Holmstedt and Cook 2018, 3).

That the Holmstedt–Abbegg database is grounded in Chomsky's generative approach is visible, among other things, in the inclusion of so-called null constituents. Because of the

[14] For other databases, which are not yet available publicly, such as Richter's database, see Kroeze (2013).

generative principle that every phrase has a 'head', a null marker has been inserted in every phrase that lacks an overt head. That Holmstedt et al. were not bound by the generative approach is visible, for example, in their non-binary hierarchical clause analysis, thus differing from Chomsky's minimalist syntax (as well as the Government-and-Binding model), which adopts a strictly binary approach to constituent structure (Holmstedt and Cook 2018, 10).

## 2.4. Back to the Black Box?

The emergence of author recognition techniques, neural networks, Artificial Intelligence, and machine learning in recent years provided new potential for biblical studies, but it also posed new challenges. The results are astonishing, but exactly what the algorithms do takes place in an impenetrable 'black box'. (The reality of machine learning is that the computer pieces together a set of patterns increasingly sophisticated until they fit the starter data, and then these patterns are used to interpret new texts. There is no known way to describe or articulate these patterns, however, which is why machine learning algorithms are spoken of as a 'black box'.) This seems to be a development that is the reverse of the openness and traceability that the 'Bible and Computer' pioneers stood for. The attempts to make Artificial Intelligence understandable to humans in the field of 'explainable

AI' (for example, in the DIANNA project [Deep Insight in Neural Network Analysis]) is still in its infancy.[15]

These transformations are perhaps even more drastic than those of the 1970s. That early period from the 1970s showed, to some extent, a continuation of pre-digital scholarly practices. For example, it became possible to look up words with a search query in a digital text file instead of a paper concordance, but that did not imply any methodological innovation.

In recent years, there have been various attempts to integrate these new developments into biblical studies by making use of advanced statistical analysis (Naaijer 2020; cf. below, §3.2), topic modelling (Vlaardingerbroek 2017), Markov Chains (Kingham et al. 2018), stylometrics (Van Hecke 2018; Van Hecke and De Joode 2021), and neural networks (Van der Schans et al. 2020; Naaijer 2020, 149–75). Here the main challenge is to determine how the results of the 'black box' relate to current scholarship.[16]

A case in point are the projects in which text clustering methods are applied to questions related to linguistic dating to see whether we can distinguish certain groups or collections of

---

[15] See 'Deep Insight And Neural Network Analysis—DIANNA', https://www.esciencecenter.nl/projects/deep-insight-and-neural-networks-analysis-dianna/, accessed 25 May 2023.

[16] Most examples in this section are taken from the ETCBC, because the present author is most acquainted with it, but the situation with other institutions and with individual researchers seems to be similar. Scholars recognise the great potential of recent developments in computer science but are still in an experimenting phase to find out how it can be made useful to biblical studies.

texts that agree with current scholarly notions such as Standard Biblical Hebrew versus Late Biblical Hebrew. The challenge is, if the outcomes agree with current scholarship, the computational analysis does not really add to our knowledge, except for confirming existing theories. But if the outcome seems to be at odds with current scholarship, should we search for explanations that still fit the traditional framework (e.g., labelling outliers in an alleged early corpus as later additions), or should we rather challenge and tweak the algorithms? And if we improve the algorithms so that they better yield the expected results, how do we avoid the risk of creating a circular argument?[17]

For all the layers of linguistic analysis that were explored with rule-based approaches and distributional analysis from the 1970s onwards, these new approaches have the potential to accelerate, refine, or automate analytical procedures. Although there are now various databases containing a morphological analysis of the Hebrew Bible, when extending the corpus to other Hebrew texts or corpora of other Northwest-Semitic languages, machine learning can be used to accelerate the process of the morphological analysis.[18] Likewise, with the search for phrase patterns, new methods searching for patterns using n-grams, flex

[17] Cf. below, §3.3, for the example of the distinction between P and non-P by author-clustering algorithms.

[18] Thus, the eScience Center project 'Morphological Parser for Inflectional Languages Using Deep Learning' aims to accelerate the analytical procedures by having the computer make more accurate predictions about the morphological analysis based on the ETCBC's existing Hebrew- and Syriac-encoded texts (Naaijer and Van Peursen 2022).

grams, etc. can replace the pattern-matching tools that functioned in the distributional analysis with which the ETCBC started (or the manual assignment of phrase patterns based on human intuition in other projects).[19] In the text-syntactic analysis, automatic anaphora resolution can complement existing methods of computer-assisted, text-hierarchical analysis based on clause relations (cf. Erwich 2021). The identification of participants is the first step to establishing their relationships as the basis for social network analysis, and other emerging approaches in which computational linguistics and literary analysis meet (cf. Canu Højgaard 2021).

## 2.5. From Talstra to Tiberias

The projects and experiments described in §2.4 show a difference from the various approaches in the early years of 'Bible and Computer'. In the projects of Andersen, Forbes, and Talstra, the lucidity of the rules that were applied served as an argument for the validity of the analysis. In those new approaches, the proof for the validity is not so much the structure of the algorithms or the analytical steps, but rather the results of test cases.[20] In, for example, the author-clustering tools that are used for Tiberias (above, §1.5), what counts as convincing argument for the analysis is the

[19] This happens in the CLARIAH Fellowship project 'PaTraCoSy: PAtterns in TRAnslation: Using COlibriCore for the Hebrew Bible corpus and its SYriac translation' (Coeckelbergs 2022).

[20] In the case of the Tiberias Stylistic Classifier (see the following discussion), at the moment of this writing the algorithms used are not publicly available.

results of a test set.[21] In their case they point to the successful deconstruction of an artificially mixed book, consisting of randomly merged segments from Jeremiah and Ezekiel, coined 'Jeriel'. The algorithms successfully distinguished between the two components of this artificial book with an accuracy of 89 to 95 percent (Dershowitz et al. 2015).

In conclusion, the potential of the machine learning algorithms is unprecedented, and the results are impressive. However, what exactly those algorithms do, and how they arrive at their results, is beyond human understanding. The insightful and traceable analyses that were the showpiece of emerging computational Bible research (above, §2.1) are now giving way to a black box that, while yielding great results, allows little insight into what goes on inside that box. Even if the output of the algorithms provides some insights (e.g., the Tiberias programs list the phenomena on which the results are based, such as typical linguistic elements of a selected corpus, which distinguish it from another corpus), the human researchers will have to find out by themselves the typical linguistic or stylistic features of a certain corpus or collection (cf. below, §3.3).

[21] The term 'author clustering' is in this context more precise than 'author recognition'; cf. Dershowitz et al. (2015, 255).

## 3.0. Use and Contributions in Biblical Studies to Date

### 3.1. Orthography

After the emergence of databases of the Hebrew Bible, it soon became clear that computational analysis enabled types of research that were hardly imaginable without digital tools. As early as the 1980s, Francis Andersen and Dean Forbes (1986) published their monumental work on spelling in the Hebrew Bible, filled with tables and mathematical formulas to investigate the distribution of *matres lectionis* over the entire biblical corpus. They could make observations about the extent to which these vowel letters were used in the biblical corpus, about the relation of the Masoretic Text to the more defective pre-exilic inscriptions and the more *plene* spellings of the last centuries BC, and about differences between the various parts of the Hebrew Bible, with the Pentateuch having the most conservative spelling. More recently, Johan de Joode and Dirk Speelman (2020) have applied quantitative linguistic methods to the orthographic heterogeneity within the Hebrew Bible and the Dead Sea Scrolls.

### 3.2. Syntax

When syntactic databases became available, all kinds of research questions could be addressed more effectively, ranging from major questions about diachronic developments (Siebesma-Mannens 2014), to the extent to which poetic structure affects clause patterns (Bosman 2019), to corresponding phrase and clause patterns in the Hebrew Bible and the Peshitta (Van

Peursen 2007; Dyk and Van Keulen 2013), and to the interpretation of specific grammatical phenomena or translation issues. It is now easy to find parallels for the construction in the phrase וַתָּשַׁר דְּבוֹרָה וּבָרָק בֶּן־אֲבִינֹעַם בַּיּוֹם הַהוּא 'On that day Deborah and Barak son of Abinoam sang (f. sg.)' (Judg. 5.1), where the verb preceding the compound subject agrees with the first element of this subject. Those parallels show that this is a common phenomenon and that an emendation of the verb or the deletion of the second part of the subject ('and Barak…') is not needed (Sandborg-Petersen 2011; Meeuse 2021, 10). Likewise, a careful analysis of the verb valence pattern used shows that the phrase וַיָּשֶׂם יְהוָה לְקַיִן אוֹת (Gen. 4.15) should be translated 'And the LORD set a sign in place on behalf of Cain' rather than with 'And the LORD put a mark on Cain', which is the rendering of the NRSV and many other translations (Dyk, Glanz, and Oosting 2013, 30–32; Meeuse 2021, 6–7).

The more advanced applications of statistical analysis and machine learning enable new possibilities for charting the distribution of clause patterns over the Hebrew Bible in relation to various parameters, such as assumed date of origin, genre, text type, and sentence pattern. An interesting case concerns the distribution of 'to be' constructions. In Biblical Hebrew, there are five ways in which 'to be' can be expressed: Bipartite and tripartite nominal clauses; constructions with the particles יֵשׁ ('there is') and אֵין ('there is not'); and clauses containing the verb הָיָה ('to be'). On the basis of quantitative analysis taking all these parameters into account, Martijn Naaijer (2020) has convincingly argued that in the alleged Early Biblical Hebrew corpus the so-called

narrative text type and the direct speech sections differ considerably, and that the direct speech sections show similarities with the Late Biblical Hebrew texts (regardless of the distinction between narrative and direct speech in the latter). In other words: in late texts, there is less of a difference between narrative and direct speech.

### 3.3. Author Clustering

Perhaps the most cutting-edge application of machine learning and computational linguistics to biblical studies can be found at the research group at Bar Ilan University that is responsible for the Tiberias Stylistic Classifier for the Hebrew Bible (above, §§1.5., 2.4., and 2.5.). Their tools distinguished between Priestly (P) and non-Priestly (non-P) texts in the Pentateuch, thus agreeing with a major conclusion of the Documentary Hypothesis (Dershowitz et al. 2015).[22] This is not only a milestone in the application of computational linguistics in biblical studies. It also shows where the interaction between computational linguistics and biblical scholarship can now take place, because the outcome of the computational analysis is not merely an 'objective proof' of a scholarly hypothesis, but rather the start of new scholarly reflection as articulated and tested with the iterative development and application of computer algorithms. Questions that arise are: How can we account for the few verses that have been classified as non-P in traditional scholarship, but were assigned

[22] Another interesting case study is the assignment of Isaiah 34–35, for which it has been argued that these chapters were written by Deutero-Isaiah (Berman 2021).

P in the computational analysis and vice versa? Do they reveal flaws in the algorithms, or should we rather reconsider their assignment to P or non-P? (For this dilemma see also above, §2.4.) What does the outcome tell about the other elements of the Documentary Hypothesis, such as the J, E, and D sources, that cannot be distinguished by the algorithms?[23] Would scholars ever have set out to answer such a question with computational linguistics if the hypothesis had not already existed? The algorithm's ability to distinguish between P and non-P means they consistently differ, but what confidence do we have that they differ in the way scholars have claimed they do (e.g. in terms of authorship and date)?

Another question relates to the notion of author recognition and computer programs that are built to detect unconscious individual elements of language use and an author's "subtle stylistic preferences" (Dershowitz et al. 2015, 253). How can such an approach be applied to compositions such as P that in Old Testament scholarship are usually considered the work of groups of scribes, or as consisting of successive editorial layers?[24] The algorithms will not reproduce S. R. Driver's (1913, 131–35) list of words and phrases typical of P, and the notion of an author as an individual that can be identified on the basis of unconscious

[23] Cf. Dershowitz et al. (2015, 270): "There appear to be two possible explanations for this: (1) the J and E source are not sufficiently distinct from one another in terms of word usage (…); (2) the traditional J/E division is flawed."

[24] See, e.g., Smend (1978, 57), on the supposed successive stages of the composition of P.

authorial fingerprints seems to be remote from the priestly circles like the alleged *Sitz im Leben* of P in traditional Old Testament scholarship. Yet, traditional source criticism and cutting-edge author-clustering algorithms largely arrive at a similar distinction between P and non-P. Here is both the requirement and opportunity to reconcile the claims of what has been called 'algorithmic criticism' (Verhaar 2016) with those of traditional scholarship. Or in other words, to proceed from Humanities 1.0 (traditional source criticism) through Humanities 2.0 (source detection with author-clustering algorithms) to Humanities 3.0 (cf. above, §1.4).

## 4.0. Prospects for Further Study, Application, and Collaboration

### 4.1. Syntax and Semantics

Most of the database projects that began in the 1970s and the 1980s started with syntax. The Andersen–Forbes database also includes semantic roles, but the way in which the labels have been assigned is not always clear and hence they are difficult to reproduce (and therefore assess). The other databases currently available also have a strong focus on syntax. This focus is understandable from the positivist climate in which these projects originated and the uneasiness that was felt with contemporaneous etymologising lexicographical approaches. But now, about half a century later, it is crucial to investigate how computational linguistics can be applied to the semantics of Biblical Hebrew. Otherwise, what happens is that advanced syntactic databases are

enriched with digital representatives of the scholarly knowledge of the nineteenth and twentieth centuries as it is codified, for example, in Brown, Driver, and Briggs's 1910 lexicon. This is what we see happen in the commercial or semi-commercial Bible software packages in which both the advanced syntactic databases discussed in this chapter and the older lexicographical resources have become available.

There are two clear ways in which the current syntactic databases could be extended towards semantic analysis. The first relates to the intersection of syntax and semantics. The search for valence patterns provides new insights about the meaning and usage of a verb. Hence one way to proceed is to enrich the syntactic labels with verbal valence patterns and the associated semantic roles according to strict criteria of how valence patterns and meaning interrelate (cf. Dyk 2016). For the study of verbal valence and clause patterns, the application of existing approaches, especially those that have been applied successfully in computational linguistics such as Role and Reference Grammar, appears to be promising (cf. Canu Højgaard 2019).

Another way in which the current database could be extended to semantic analysis is the application of methods that have been developed in computational lexicography and semantics to the Hebrew Bible. Obviously, not everything that has been developed in this field is applicable to the Bible, which is, linguistically speaking, a limited corpus without native speakers. Thus, building a WordNet for Biblical Hebrew would meet with many complications. What could be promising, however, is to experiment with approaches such as co-occurrence analysis,

topic modelling, and similar methods to establish the relations between words.[25]

## 4.2. Linked Data and Geospatial Analysis

An extension of semantic and lexicographic information may be the interlinking with other resources. In recent years, the ETCBC has explored the potential use of Linked Data in which textual data is linked to encyclopedic or geospatial data. Pilot projects include Linking Syriac Data (2017–2018);[26] Linking Syriac Liturgies (Van Peursen and Veldman 2018); and Linking Syriac Geographic Data (see Van Peursen 2018). Although it is wonderful that this brings the textual data (in these projects: Syriac data) into the Linked Data universe, there is the danger that encyclopedic information takes the place of sound syntactic analysis. If, for example, we want to map all the geographical entities mentioned in the Syriac Book of the Laws of the Countries, we have to decide how to identify the places mentioned or to locate the peoples mentioned in those texts. The same can be said of the famous catalogue of nations and peoples gathered at Jerusalem in Acts 2:9–11 (Van Altena 2022, 135–57). Such questions may

[25] Such new initiatives could be linked with or even incorporated in the most up-to-date digitally-available lexicographical and semantic resources, such as those of the Semantic Dictionary of Biblical Hebrew (https://semanticdictionary.org, accessed 4 May 2023) and the Semantics of Ancient Hebrew Database (https://www.sahd.div.ed.ac.uk, accessed 4 May 2023) projects.

[26] See https://github.com/hvlaardingerbroek/LinkSyr, accessed 4 May 2023.

be even more challenging in the case of biblical studies, because of the debate over the extent to which the biblical accounts can be related to the history and geography of ancient Israel and given the uncertainties about the identification of places and events in the Bible.

In addition to these interpretive difficulties, there is the challenge that we mentioned in §1.3 above. If linked geospatial data do not go beyond a mere digital representation of the well-known traditional atlases of the Bible, or of the maps of Jerusalem, ancient Israel, the ancient Near East, and the Roman Empire often included in printed Bibles, this only serves practical purposes, rather than representing a methodological innovation. However, given the 'spatial turn' in biblical studies (cf. Van Altena 2022, 41), it is to be expected that geospatial analysis, when applied properly, can lead to new insights and a better understanding of the Bible, even though its application to the Bible is still in its infancy.

## 4.3. Collaboration and Open Science

Another field where progress can be made is the comparison of the various linguistic databases of the Hebrew Bible. Each database has its specific approach, and the user is most helped by being able to compare the various databases, their underlying assumptions, and the way in which these assumptions resulted in the annotations in each verse of the Bible.[27] It is a pity, however, that anyone who wants to compare the three most elaborate

[27] A nice comparison is made in Miller-Naudé and Naudé (2018).

databases available to date (cf. §§1.1 and 2.3) needs to purchase them in commercial software packages. Only the ETCBC database and the Westminster Hebrew Syntax Without Morphology (cf. n. 6 above) are publicly available. Bringing the other databases into the open access domain is easier said than done, given copyright issues and business models, as well as practical challenges, but hopefully these challenges can be resolved in the near future. This will be necessary to enable scholarly pursuits engaging with all the databases.

Open Science, however, is more than making databases available. It relates also to the transparency of analytical procedures and the availability of queries and algorithms. A breakthrough in the application of computational linguistic analysis to the Hebrew Bible would be the availability of the workflows of the data creation processes and the programs that have been used in the creation of those databases and of the algorithms that are currently being developed for advanced cutting-edge approaches as those mentioned in §§2.4. and 3.3.

## 4.4. Computational Linguistics and the New Testament

This chapter focused on the use of computational linguistics in Old Testament studies. In New Testament studies we see parallel developments, although syntactic databases emerged somewhat later than in Old Testament studies.[28]

[28] For the pioneering work in the 1970s and 1980s see Mealand (1988). However, in the first decades of the emerging field of 'Bible and Computer' relatively more attention was paid to the Old Testament and

The morphological encoding started, as in the case of the Old Testament, in the 1970s, with, e.g., the GRAMCORD Greek New Testament (first published 1977); the work of Timothy and Barbara Friberg, who produced the Analytical Greek New Testament (first published 1981); and MorphGNT, which was initiated in the 1980s by Robert Kraft at the University of Pennsylvania's Center for Computer Analysis of Texts (CCAT) and received major updates and corrections by James Tauber from the 1990s onwards.[29]

The computational syntactic analysis of the New Testament received an impetus from two projects that started in the first years of the twenty-first century.[30] The first project is the Greek

---

Hebrew than to the New Testament and Greek. This is reflected, for example, in the contributions to the AIBI conferences (cf. above, §1.2). The contributions to the first AIBI conference (Leuven, 1985) included twelve contributions that dealt exclusively with Hebrew and the Old Testament and only four that dealt with Greek and the New Testament (besides eight other contributions). The second conference (Jerusalem, 1988) showed similar statistics. It contained seventeen contributions on Hebrew and the Old Testament, two on the Septuagint, two on Greek and the New Testament and one on the Greek works of Gregorius of Nyssa (besides thirteen other on general issues or discussing both the Old and the New Testament).

[29] Available on GitHub: https://github.com/morphgnt, accessed 4 May 2023.

[30] Because of copyright issues, these open-source projects are often based on the older editions by Nestle, Tischendorf and Westscott, and Hort, or on the SBL Greek New Testament, rather than on the most recent Nestle-Aland edition. For a morphologically annotated version of

New Testament of the OpenText.org initiative by Stanley E. Porter at McMaster Divinity College and partners. The goal of this project is, according to its website, "to construct a representative corpus of Hellenistic Greek (including the entire New Testament and selected Hellenistic writings of the same period) to facilitate linguistic and literary research of the New Testament documents." At clause level their annotations include four major categories: Subject, Predicator, Complement, and Adjunct.

Some more syntactic categories (e.g., object, second object) are distinguished in a project of the Asia Bible Society, namely the Greek syntax trees produced by Andi Wu and Randall K. Tan (who was also involved in the OpenText.org project) and made available through Clear Bible (formerly Global Bible Initiative).[31] These data interact well with other tools such as the Lowfat Syntax Tree Browser.[32]

---

the Byzantine Text see https://github.com/byztxt, accessed 4 May 2023.

[31] Greek syntax trees: https://github.com/biblicalhumanities/greek-new-testament/tree/master/syntax-trees, accessed 4 May 2023; Clear Bible: https://www.clear.bible, accessed 4 May 2023.

[32] See https://github.com/biblicalhumanities/greek-new-testament/tree/master/syntax-trees/reader/doc, accessed 4 May 2023. For a newer release see https://github.com/Clear-Bible/macula-greek, accessed 4 May 2023.

A project to bring the data from the OpenText project and those from the Asia Bible Society together in Text-Fabric is carried out by Oliver Glanz at the Center of Biblical Languages and Computing (CBLC) at Andrews University.[33]

Whereas computational Old Testament studies had a strong linguistic focus from the 1970s onwards, in New Testament studies there were other areas in which the potential of the computer was explored first, such text editing, stemmatology, and manuscript studies. The computer program Collate, developed by Peter Robinson in the late 1980s (succeeded in 2010 by Collatex[34]) was soon adopted by New Testament scholarship in Birmingham and Münster for text comparison and text editing. In the early 1990s, Gerd Mink, one of the editors of the *Editio Critica Maior* (ECM) of the Greek New Testament, developed the Coherence-Based Genealogical Method (CBGM; Wachtel 2019). This method was particularly apt to deal with the typical features of the transmission of the New Testament, such as the high degree of contamination, which hinders the traditional genealogical tree-model (Gurry 2016).

New Testament scholarship has also made great progress in manuscript imaging (see various contributions in Hamidović et al. 2019). Hundreds of manuscripts have been digitised and high-quality images can be studied and annotated in the Virtual Manuscript Room of the *Institut für Neutestamentliche Textforschung*

[33] See https://github.com/CenterBLC/NA, accessed 4 May 2023.

[34] See https://collatex.net/about, accessed 4 May 2023.

(INTF) in Münster.[35] And as in the case of Old Testament studies, new experiments with data meaning, text reuse detection, and the use of NLP for semantic information extraction have appeared on the scene (for examples see Hamidović et al. 2019).[36]

## 5.0. Further Reading

The various databases discussed in this chapter do not provide final answers, but are useful tools, each of them situated in the complex field of linguistic theories. It is therefore extremely important to use them in consultation with the documentation listed below.

Those who want to do more advanced analysis are advised to develop some basic programming skills and use the datasets that are available as a whole on GitHub or another platform, rather than only with a user-friendly search interface.

In the case of the ETCBC database, for example, Meeuse (2021) is a good starting point for exploring the database through the user interface of the SHEBANQ website, but much more advanced research (as in the examples mentioned in §3.2) is possible for those who have mastered Python and use the Python package Text-Fabric to analyse the Hebrew Bible.[37]

---

[35] See https://ntvmr.uni-muenster.de, accessed 4 May 2023.

[36] See also above, §4.2 on geospatial analysis in New Testament studies.

[37] SHEBANQ website: https://shebanq.ancient-data.org; ETCBC database on GitHub: https://github.com/ETCBC/bhsa; Text-Fabric: https://github.com/annotation/text-fabric; Python courses: https://www.codecademy.com or https://www.udemy.com/user/fredbaptiste. All accessed 4 May 2023.

## 5.1. Computational Linguistics

1. Clark, Fox, and Lappin (2010)
2. Jurafsky and Martin (2021)

## 5.2. Hermeneutical Implications of Computational Text Analysis

1. Bod (2013)
2. Clivaz (2019)
3. Van der Weel (2011)
4. Van Peursen (2010)

## 5.3. History of the Discipline

1. Oosting (2016)
2. Poswick (2010)

## 5.4. Database and Tools

### 5.4.1. General Overview and Methodological Issues

1. Kroeze (2013)
2. Miller-Naudé and Miller (2018)

### 5.4.2. Andersen-Forbes Database

1. Andersen and Forbes (2012)

### 5.4.3. Accordance Syntactic Database

1. Accordance documentation (2014)
2. Holmstedt and Cook (2018)

### 5.4.4. ETCBC Database

1. Kingham (2018)
2. Kingham and Van Peursen (2018)
3. Meeuse (2021)

### 5.4.5. Tiberias Stylistic Classifier

1. Berman (2021)
2. Dershowitz et al. (2015)

### 5.4.6. New Testament Databases

1. Porter et al. (2019)

## References

Accordance documentation. 2014. 'Holmstedt–Abegg Hebrew Syntactic Database. Principle and Parameters, v. 5.0 (rev. October 2014)'. https://www.accordancefiles1.com/exchange/downloads/documents/holmstedt_syntax_14.pdf, accessed 23 May 2022.

action.ai. 2021. 'Natural Langauge is an Unruly Beast: How We Tame Her in Order to Create Groundbreaking Conversational AI'. https://action.ai/what-happens-when-we-speak-a-brief-foray-into-language-and-how-we-artificially-process-it/, accessed 29 June 2022.

Andersen, Francis I. and A. Dean Forbes. 1986. *Spelling in the Hebrew Bible*. Biblica et Orientalia 41. Rome: Biblical Institute Press.

———. 2012. *Biblical Hebrew Grammar Visualized*. Linguistic Studies in Ancient West Semitic 6. Winona Lake, IN: Eisenbrauns.

Barr, James. 1961. *The Semantics of Biblical Language*. Oxford: Oxford University Press.

Berman, Joshua. 2021. 'Measuring Style in Isaiah: Isaiah 34–35 and the Tiberias Stylistic Classifier for the Hebrew Bible'. *Vetus Testamentum* 71 (3): 303–16. doi.org/10.1163/15685330-12341070.

Bod, Rens. 2013. 'Who's Afraid of Patterns? The Particular versus the Universal Meaning of Humanities 3.0'. *BMGN Low Countries Historical Review* 128: 171–9.

Bosman, Hendrik-Jan. 2019. 'Prosodic Influence on the Text Syntax of Lamentations'. PhD dissertation, Vrije Universiteit Amsterdam.

Canu Højgaard, Christian. 2019. 'Semantic Mapping of Participants in Legal Discourse'. *HIPHIL Novum* 5 (2): 136–42.

———. 2021. 'Roles and Relations in Biblical Law: A Study of Participant Tracking, Semantic Roles, and Social Networks in Leviticus 17–26'. PhD dissertation, Vrije Universiteit Amsterdam.

Clark, Alexander, Chris Fox, and Shalom Lappin. 2010. *The Handbook of Computational Linguistics and Natural Language Processing*. Malden, MA: Wiley-Blackwell.

Clivaz, Claire. 2019. *Ecritures digitales: Digital Writing, Digital Scriptures*. Digital Biblica Studies 4. Leiden: Brill.

Coeckelbergs, Mathias. 2022. 'From Pattern to Interpretation. Using Colibri Core to Detect Translation Patterns in the

Peshitta'. In *LREC 2022 Conference Proceedings*, edited by Nicoletta Calzolari et al., 4270–74. Paris: ELRA, 2022.

de Groot, A. D. 1946. *Het denken van den schaker: Een experimenteel-psychologische studie*. Amsterdam: Noord-Hollandsche Uitgevers Maatschappij.

de Joode, Johan, and Dirk Speelman. 2020. 'A Hermeneutic of Variation? The Orthographic Variability of the Hebrew Bible and the Larger Dead Sea Scrolls'. *Journal for Semitics* 29 (2) [24 pages]. doi.org/10.25159/2663-6573/6633.

Dershowitz, Idan, Navot Akiva, Moshe Koppel, and Nachum Dershowitz. 2015. 'Computer Source Criticism of Biblical Texts'. *Journal of Biblical Literature* 134 (2): 253–71. doi.org/10.15699/jbl.1342.2015.2754.

Dorobantu, Marius. 2022. 'Imago Dei in the Age of Artificial Intelligence: Challenges and Opportunities for a Science-Engaged Theology'. *Christian Perspectives on Science and Technology*, n.s., 1: 175–96. doi.org/10.58913/KWUU3009.

Driver, Samuel Rolles. 1913. *An Introduction to the Literature of the Old Testament*. 9th rev. ed. Edinburgh: T&T Clark.

Dyk, Janet W. 2016. 'How do Hebrew Verbs Differ? A Flow Chart of Differences'. In *Contemporary Examinations of Classical Languages: Valency, Lexicography, Grammar*, edited by Timothy Martin Lewis, Alison G. Salvesen, and Beryl Turner, 33–51. Perspectives on Linguistics and Ancient Languages 8. Piscataway, NJ: Gorgias.

Dyk, Janet, Oliver Glanz, and Reinoud Oosting. 2013. 'Het belang van valentiepatronen voor het vertalen van Bijbels

Hebreeuwse werkwoorden'. *Met Andere Woorden* 32 (2): 23–35.

Dyk, Janet W., and Percy S. F. van Keulen. 2013. *Language System, Translation Technique, and Textual Tradition in the Peshitta of Kings*. Monographs of the Peshitta Institute Leiden 19. Leiden: Brill.

Erwich, Christiaan M. 2021. 'Who is Who in the Psalms? A Computational Analysis of Participants and Their Networks'. PhD dissertation, Vrije Universiteit Amsterdam.

Gurry, Peter J. 2016. 'How Your Greek NT is Changing: A Simple Introduction to the Coherence-Based Genealogical Method (CBGM)'. *Journal of the Evangelical Theological Society* 59 (4): 675–89.

Hamidović, David, Claire Clivaz, and Sarah Bowen Savant (eds). 2019. *Ancient Manuscripts in Digital Culture: Visualisation, Data Mining, Communication*. Digital Biblical Studies 3. Leiden: Brill, 2019.

Hardmeier, Christof Felix. 1970. 'Die Verwendung von elektronische Datenverarbeitungsanlagen in die alttestamentliche Wissenschaft: Neue Möglichkeiten der Forschung am Alten Testament'. *Zeitschrift für die alttestamentliche Wissenschaft* 83: 175–85.

Holmstedt, Robert D., and John A. Cook. 2018. 'The Accordance Hebrew Syntactic Database Project'. *Journal of Semitics* 27 (1) [23 pages]. doi.org/10.25159/1013-8471/3010.

Jurafsky, Dan, and James H. Martin. 2021. *Speech and Language Processing*. 3rd ed. Draft. https://web.stanford.edu/~jurafsky/slp3/, accessed 4 May 2023.

Kaajan, Marianne. 2019. 'Is Biblical Hebrew a Non-Configurational Language? Reconsidering the Evidence from Discontinuous Phrases'. *HIPHIL Novum* 5 (2): 45–69.

Kingham, Cody. 2018. 'Data Creation (Updated version 03.04.18)'. http://www.etcbc.nl/datacreation/, accessed 23 May 2023.

Kingham, Cody, Etienne van de Bijl, Sandjai Bhulai, and Wido van Peursen. 2018. 'A Probabilistic Approach to Linguistic Variation and Change in Biblical Hebrew'. https://github.com/ETCBC/Probabilistic_Language_Change, accessed May 23, 2022.

Kingham, Cody, and Wido van Peursen. 2018. 'The ETCBC Database of the Hebrew Bible'. *Journal for Semitics* 27 (1) [13 pages]. doi.org/10.25159/1013-8471/2974.

Kittel, Gerhard, and Gerhard Friedrich (eds). 1933–79. *Theologisches Wörterbuch zum Neuen Testament*. Stuttgart: Kohlhammer.

Kroeze, Jan H. 2013. 'Computational Information Systems: Biblical Hebrew'. In *Encyclopedia of Hebrew Language and Linguistics* 1: 527–34, edited by Geoffrey Khan. Leiden: Brill.

Longacre, Robert E. 1989. *Joseph: A Story of Divine Providence*. Winona Lake, IN: Eisenbrauns.

Mealand, David. 1988. 'Computers in New Testament Research: An Interim Report'. *Journal for the Study of the New Testament* 33 (2): 97–115.

Meeuse, Bas. 2021. 'SHEBANQ Tutorial 2021. How to Start Using the BSHA Database'. https://github.com/ETCBC/Tutorials/blob/master/SHEBANQ%20tutorial%202021.pdf, accessed 23 May, 2022.

Miller-Naudé, Cynthia L., and Jacobus A. Naudé. 2018. 'New Directions in the Computational Analysis of Biblical Hebrew Grammar'. *Journal for Semitics* 27 (1) [17 pages]. doi.org/10.25159/1013-8471/4628.

Naaijer, Martijn. 2020. 'Syntactic Variation in Clause Structure in Biblical Hebrew'. PhD dissertation, Vrije Universiteit Amsterdam.

Naaijer, Martijn, and Wido van Peursen 2022. 'Parsing Hebrew and Syriac Morphology using Deep Learning: State-of-the-art Technology Meets Ancient Literature'. https://blog.esciencecenter.nl/parsing-hebrew-and-syriac-morphology-using-deep-learning-cb6832bb6685, accessed 23 May, 2022.

Oosting, Reinoud. 2016. 'Computer-Assisted Analysis of Old Testament Texts: The Contribution of the WIVU to Old Testament Scholarship'. In *The Present State of Old Testament Studies in the Low Countries: A Collection of Old Testament Studies Published on the Occasion of the Seventy-Fifth Anniversary of the Oudtesta-mentisch Werkgezelschap*, edited by Klaas Spronk, 192–209. Oudtestamentische Studiën 69. Leiden: Brill.

Porter, Stanley E., Christopher D. Land, and Francis G. H. Pang. 2019. *Linguistics and the Bible: Retrospects and Prospects*. McMaster New Testament Studies 9. Eugene, OR: Pickwick Publications.

Postma, Ferenc, Eep Talstra, and Marc Vervenne. 1983. *Exodus: Materials in Automatic Text Processing*. Amsterdam: VU Boekhandel.

Poswick, R. Ferdinand. 2010. 'From Louvain-la-Neuve (1985) to El Escorial in Madrid (2008): 25 Years of AIBI'. In *Computer Assisted Research on the Bible in the 21st Century*, edited by Luis Vegas Montaner, Guadalupe Seijas de los Ríos-Zarzosa, and Javier del Barco, 3–11. Bible in Technology 3. Piscataway, NJ: Gorgias.

Roorda, Dirk. 2018. 'Coding the Hebrew Bible'. *Research Data Journal for the Humanities and Social Sciences* 3 (1): 27–41. doi.org/10.1163/24523666-01000011.

Sanborg-Petersen Ulrik. 2011. 'On Biblical Hebrew and Computer Science: Inspiration, Models, Tools, and Cross-Fertilization'. In *Tradition and Innovation in Biblical Interpretation: Studies Presented to Professor Eep Talstra on the Occasion of his Sixty-Fifth Birthday*, edited by Willem Th. van Peursen and Janet W. Dyk, 261–76. Studia Semitica Neerlandica 57. Leiden: Brill.

Siebesma-Mannens, Femke. 2014. 'Continuity and Discontinuity. A Study in Biblical Hebrew on the Variation of the Prepositions אֶל and לְ Occurring with the Verb אמר'. MA thesis, Vrije Universiteit Amsterdam.

Smend, Rudolf. 1978. *Die Entstehung des Alten Testaments*. Stuttgart: Kohlhammer.

Talstra, Eep. 2010. 'In the Beginning, when Making Copies used to Be an Art…: The Bible among Poets and Engineers'. In *Text Comparison and Digital Creativity: The Production of Presence and Meaning in Digital Text Scholarship*, edited by Wido van Peursen, Ernst D. Thoutenhoofd, and Adriaan van der Weel, 31–56. Scholarly Communication 1. Leiden: Brill.

Talstra, Eep and Janet W. Dyk. 2006. 'The Computer and Biblical Research: Are there Perspectives beyond the Imitation of Classical Instruments?'. In *Text, Translation, and Tradition: Studies on the Peshitta and Its Use in the Syriac Tradition Presented to Konrad D. Jenner on the Occasion of his Sixty-Fifth Birthday*, edited by W. Th. van Peursen and R. B. ter Haar Romeny, 189–203. Monographs of the Peshitta Institute Leiden 14. Leiden: Brill.

van Altena, Vincent Paul. 2022. 'What has Athens to Do with Jerusalem: The Potential of Spatial-Temporal Analysis Methods to Interpret Early Christian Literature'. PhD dissertation, Technical University Delft.

van der Schans, Yanniek, David Ruhe, Wido van Peursen, and Sandjai Bhulai. 2020. 'Clustering Biblical Texts Using Recurrent Neural Networks'. In *Proceedings of the Network Institute Academy Assistants program 2018/2019*, edited by Victor de Boer, Antske Fokkens, Christine Moser, and Ivar Vermeulen. doi.org/10.5281/zenodo.4003509.

van der Weel, Adriaan. 2011. *Changing our Textual Minds: Towards a Digital Order of Knowledge*. Manchester: Manchester University Press.

Van Hecke, Pierre. 2018. 'Computational Stylometric Approach to the Dead Sea Scrolls: Towards a New Research Agenda'. *Dead Sea Discoveries* 25 (1): 57–82.

Van Hecke, Pierre, and Johan de Joode. 2021. 'Promises and Challenges in Designing Stylometric Analyses for Classical Hebrew'. In *Hebrew Texts and Language of the Second Temple Period: Proceedings of an Eighth Symposium on the Hebrew of*

*the Dead Sea Scrolls and Ben Sira*, edited by Steven Fassberg, 349–74. Studies on the Texts of the Desert of Judah 134. Leiden: Brill.

van Peursen, Willem Th. 2007. *Language and Interpretation in the Syriac Text of Ben Sira: A Comparative Linguistic and Literary Study*. Monographs of the Peshitta Institute Leiden 16. Leiden: Brill.

———. 2010. 'Text Comparison and Digital Creativity: An Introduction'. In *Text Comparison and Digital Creativity: The Production of Presence and Meaning in Digital Text Scholarship*, edited by Wido van Peursen, Ernst D. Thoutenhoofd, and Adriaan van der Weel, 1–27. Scholarly Communication 1. Leiden: Brill.

———. 2015. 'Mathematical Rigour and Scholarly Intuition: Some Reflections on Andersen's and Forbes' *Biblical Hebrew Grammar Visualized*'. *Ancient Near Eastern Studies* 52: 298–307. doi.org/10.2143/ANES.52.0.3082875.

———. 2018. 'Introduction: Linking Syriac Geographic Data'. https://medium.com/pelagios/introduction-linking-syriac-geographic-data-3e7a8f88dede, accessed 4 May 2023.

———. 2020a. 'Tracing Text Types in Biblical Hebrew'. *Vetus Testamentum* 70 (1): 140–55. doi.org/10.1163/15685330-12341430.

———. 2020b. 'De computer en de Geest: Digital Humanities en het verstaan van de Bijbel'. *Radix* 46 (4): 298–308.

van Peursen, W. T., and J. G. Veldman. 2018. *Linking Syriac Liturgies*. DANS. doi.org/10.17026/dans-26t-hhv7.

Verhaar, Peter. 2016. 'Affordances and Limitations of Algorithmic Criticism'. PhD dissertation, Leiden University.

Vlaardingerbroek, Hannes. 2017. 'Do You See What I Am Talking About? Towards a Topic Visualizer for Syriac Texts' Project Report'. https://www.academia.edu/30737774/Do_you_see_what_I_am_talking_about, accessed 23 May 2022.

Wachtel, Klaus 2019. 'The Development of the Coherence-Based Genealogical Method (CBGM), its Place in Textual Scholarship, and Digital Editing'. In *The Future of New Testament Scholarship*, edited by Garrick V. Allen, 435–46. Wissenschaftliche Untersuchungen zum Neuen Testament 417. Tübingen: Mohr Siebeck.