(재)대한성서공회

「성경원문연구」 44 (2019. 4.), 237-253
ISSN 1226-5926 (print), ISSN 2586-2480 (online)
DOI: https://doi.org/10.28977/jbtr.2019.4.44.237
https://dbpiaone.com/bskorea/index.do

JBTR
JOURNAL OF BIBLICAL TEXT RESEARCH

A Computational Approach to Syntactic Diversity in the Hebrew Bible

Wido van Peursen*

1. Introduction

For more than four decades, the Eep Talstra Centre for Bible and Computer (ETCBC)[1] has been building a richly-annotated linguistic database of the Hebrew Bible. This database contains the complete text of the Hebrew Bible and annotations on how the morphemes and lexemes combine to build words (morphology), the combination of words into phrases, and phrases into clauses (syntax), and the merging of clauses into larger text structures (text level).

When Eep Talstra started this research center in the late 1970s, he was one of the pioneers in the field of Bible and Computer. It was still the time of punch cards and mainframe computers.[2] The application of a computational approach in textual analysis and biblical exegesis was still a largely unexplored area. Therefore, the history of the ETCBC not only tells the history of a discipline, but also reflects the technical developments in computing and text analysis from the first pioneering experiments up to the present.[3]

* Ph.D. in Semitic Languages, Leiden University (1999) . Professor of Old Testament at Vrije Universiteit Amsterdam, and director of the Eep Talstra Centre for Bible and Computer (ETCBC). w.t.van.peursen@vu.nl.

[1] www.etcbc.nl. The ETCBC was formerly known as WIVU, Werkgroep Informatica Vrije Universiteit, but in 2013 it was renamed after its founder, Professor Eep Talstra.

[2] See the online exhibition on the occasion of the 40th anniversary of the ETCBC at http://expo.ubvu.vu.nl/carousel_intro.html?index=0&expoid=12&lang=_nl.

[3] On the history of the ETCBC see further Reinoud Oosting, "Computer-Assisted Analysis of Old Testament Texts: The Contribution of the WIVU to Old Testament Scholarship", Klaas Spronk,

In the last few years, the ETCBC dataset, published as the BHSA (BHS + Amstelodamensis), has gone through various transformations. Until about six years ago, it was only accessible through a local server at the Vrije Universiteit Amsterdam, but since then it has been made accessible online through the SHEBANQ website,[4] whereas the complete database is also available on Github, a popular web-based hosting service for software and data.[5] For those who want to perform more advanced computational analysis of the Hebrew Bible, Dirk Roorda (DANS: Data Archiving and Networked Services) developed Text-Fabric, a powerful tool for text storage and analysis that is now available both as a browser and as a Python package.

Over the years, the database has been the basis of scholarly research into the Hebrew Bible, including the Ph.D. dissertations by Daniel Ryou (류호준), *Zephanaiah's Oracles Against the Nations: A Synchronic and Diachronic Study of the Composition of Zephanaiah 2:1-3:8* (Vrije Universiteit Amsterdam, 1994; Ph.D. supervisors: Henk Leene and Eep Talstra); Young Bok Park (박영복), *Restoration in the Book of Ezekiel: A Text-Linguistic Study of Ezekiel 33-39* (Vrije Universiteit Amsterdam, 2013; Ph.D. supervisors: Eep Talstra and Eveline van Staalduine-Sulman); and Byungduck Park (박병덕), *Praise, Apathy, and Doubt: The Narrative Intention of 2 Sam 12:16-23* (Protestant Theological University, 2014; Ph.D. supervisors: Klaas Spronk and Eveline van Staalduine-Sulman). Currently, it provides the basis for the ongoing Ph.D. research by Gyusang Jin (진규상), *Investigating the Text-hierarchical Structures and Composition of Numbers* (Ph.D. supervisors: Wido van Peursen and Joep Dubbink).

Since the database is freely available, people have built their own tools and interfaces upon it. Thus, Kyoungsik Kim (김경식) (Bar-Ilan University, Ramat Gan) built the Korean Bible study tool AlphAlef (알파알렙 온라인 성경)[6] on the basis of the ETCBC data because he "felt the necessity for the free and useful professional Bible software for the Korean readers who eagerly want to

ed., *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 Oudtestamentisch Werkgezelschap* (Leiden: Brill, 2016), 192-209.

[4] https://shebanq.ancient-data.org.

[5] https://github.com/etcbc/bhsa.

[6] http://app.alphalef.com.

read and understand the Bible in its original languages."[7] For Bible translators it may be useful that Reinier de Blois (United Bible Societies) built a tool for using the ETCBC data within Paratext, the software used in many Bible translation projects.[8] A systematic analysis of linguistic patterns has a great impact on the theory and practice of Bible translation.[9] It also provides insight in the translation practices of the ancient versions.[10] The ETCBC database has also been the basis for the Bible Online Learner,[11] a tool for learning Biblical Hebrew based on Persuasive Learning Design. This tool was developed by Nicolai Winther Nielsen (Fjellhaug International University College Denmark) and his team and has now English, Danish, Spanish, Portugese and Chinese interfaces.

In this contribution, I will focus on one aspect of Biblical Hebrew studies for which the computational approach has proven to be very helpful: the linguistic diversity in the Hebrew Bible. But first I will give a short overview of the data creation process underlying the ETCBC database of the Hebrew Bible.

2. The ETCBC data creation pipeline

The analytical procedures in the creation of the database follow a bottom-up approach. This means that the analysis starts at the level of morphemes, and from there moves to the higher linguistic levels of words, phrases, clauses, sentences, and texts. The analytical procedures also follow the form-to-function principle, which means that the distributional patterns of linguistic phenomena

[7] Kim, Kyoungsik, "ETCBC Data for the Libre Bible Software", http://etcbc.nl/uncategorized/etcbc-data-for-the-libre-bible-software (22 April 2019).

[8] It is available in Paratext 8.

[9] Many publications by researchers of the ETCBC deal implicitly or explicitly with questions of Bible translation. To give just one example, see: J. W. Dyk, "Deportation or Forgiveness in Hosea 1:6? Verb Valence Patterns and Translation Proposals", *The Bible Translator* 65:3 (2014), 235–279. See also below, section 2, on the example from the book of Joel.

[10] See, for example, the following monographs on the Peshitta that were produced in research projects of the ETCBC: W. Th. van Peursen, *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, 2007); J. W. Dyk and P. S. F. van Keulen, *Language System, Translation Technique and Textual Tradition in Peshitta Kings*, Monographs of the Peshitta Institute Leiden 19 (Leiden: Brill, 2013).

[11] https://bibleol.3bmoodle.dk/

are described first before functions are assigned. For example, at word level the morphemes are identified (e.g. a prefix, a lexeme and a suffix that constitute a verb form) before the assignment of morpho-syntactical functions (e.g. “Imperfect 3rd pers. sing.”).[12] In this respect, the creation of the ETCBC database differs from many other tagging projects, which simply tag an entire word without calculating the morphological components.

Thus, the database of the Hebrew Bible is created according to the bottom-up and form-to-function principles. First, the Hebrew and Aramaic words of the Old Testament are segmented into morphemes. From this morphological analysis, functional deductions are made. After the word level analysis, the words are combined into phrases. The phrase level analysis involves a lexicographical analysis (determination of lexical class) and morpho-syntactic analysis (the systematic adaptation of word classes in certain environments). A distinction is made between a word’s default part of speech, which can be found in the lexicon, and a phrase-dependent part of speech, that is, the part of speech that a word adopts in a certain environment. This appears to be a useful way to face the challenge that words can assume various parts of speech in different contexts, such as an adjective that may function as a noun, or a participle that may function as an adjective.

In the next step of the analytical procedures the phrases are combined into clauses. The text is segmented to decide which phrases together constitute a clause, a clause being defined as any construction with at most[13] one predicate. Subsequently, the clauses are parsed. Functions such as subject, object and adjunct are assigned. The next step concerns the relationship between clauses and proposals for discourse analysis. This higher-level analysis, where syntactic observations are used for establishing the structure and delimitation of a text, especially leads to interesting encounters with other exegetical approaches, which often delimit the text according to thematic or theological considerations. To give just one example, in Joel 2:18 the imperfect consecutive *wayqanne*’ is

[12] For the methodological and practical advantages of this approach, see, among others, my “Progress Report: Three Leiden Projects on the Syriac Text of Ben Sira”, R. Egger-Wenzel, ed., *Ben Sira’s God. Proceedings of the Second International Ben Sira Conference, Durham, Ushaw College, 2001*, Beihefte zur Zeitschrift für die Alttestamentliche Wissenschaft 321 (Berlin: De Gruyter, 2002), 361-370.

[13] There are also clauses without predication, for instance, in the case of ellipsis.

grammatically a past tense. The past interpretation fits the verb form, as well as the overall text-syntactic structure of this chapter. However, many Bible translations (including, for example, the NIV) translate with a future tense, presumably because of a thematic division of the book of Joel into judgement (1:1–2:11), repentance (2:12–17) and promise of salvation (2:18-3[4]:21), and the assumption that the last section should refer to the future. The ETCBC researchers will in such a case advocate the priority of linguistic observations over theological considerations.[14)]

The data creation process is thus an interactive effort, in which the human researchers and the computer interact to enrich the Hebrew text with the encodings at all linguistic levels. In these interactive procedures, use is made of (1) programs that recognize the patterns of formal elements which combine to form words, phrases, clauses and textual units; (2) language-specific auxiliary files like a lexicon and a description of the morphology; (3) data sets, built up gradually, so that all patterns registered in previous analyses are included; and (4) programs that use the data sets and the auxiliary files to make proposals in the interactive analysis.[15)]

The structure of the database in which all the results of the ETCBC pipeline are stored and its underlying text model received a strong methodological foundation in Crist-Jan Doedens' Ph.D. dissertation on text databases,[16)] which was further implemented in the powerful text database engine EMDROS.[17)]

[14)] Cf. Eep Talstra, "Text, Tradition, Theology: The Example of the Book of Joel", E. Van der Borght and P. van Geest, eds., *Strangers and Pilgrims on Earth: Essays in Honour of Abraham van de Beek* (Leiden: Brill, 2011), 309–327, esp. 324–325; W. T. van Peursen, "'This is What Was Spoken by the Prophet Joel'. The Latter Rain in Joel's Prophecies and in Dutch Pentecostalism", M. Klaver, S. Paas, and E. van Staalduine-Sulman, eds., *Evangelicals and Sources of Authority*, Amsterdam Studies in Theology and Religion 6 (Amsterdam: VU University Press, 2016), 271–285.

[15)] For more details see P. S. F. van Keulen and W. Th. van Peursen, eds., *Corpus Linguistics and Textual History: A Computer-Assisted Interdisciplinary Approach to the Peshitta*, Studia Semitica Neerlandica 48 (Assen: Van Gorcum, 2006); W. Th. Van Peursen, *Language and Interpretation*, Chapters 7–8; for more information about the ETCBC data creation pipeline see Cody Kingham's contribution("Data Creation") on http://www.etcbc.nl/datacreation/ (22 April 2019).

[16)] Crist-Jan Doedens, *Text Databases: One Database Model and Several Retrieval Languages*, Language and Computers 14 (Amsterdam and Atlanta: Rodopi, 1994).

[17)] See Ulrik Petersen, "Emdros – A Text Database Engine for Analyzed Or Annotated Text" (Geneva: Proceedings of the COLING Conference, 2004).
Available online: https://emdros.org/petersen-emdros-COLING-2004.pdf.

Doedens' foundational principles were extended in the SHEBANQ project with the development of LAF-fabric, which combines Doedens' principles with the Linguistic Annotation Framework. Later on, LAF-fabric was further developed into Text-Fabric, which greatly simplifies the data model by removing unnecessary overhead.[18]

This labour-intensive data creation process for the whole Hebrew Bible took almost forty years. Not only because of the huge amount of work involved in it, but also because no tools or guiding principles were available when the project started (see above, section 1). Now that the encoding of the Hebrew Bible has been completed and other corpora (Dead Sea Scrolls, Peshitta, other Hebrew and Syriac texts) appear at the horizon, the main challenge is: how can we accelerate the analytical procedures by improving and innovating our computational toolkit, while preserving the valuable principles (bottom-up, form-to-function) and insights (e.g. about the phrase dependent part of speech) that have guided us in the past four decades.

3. Linguistic diversity in the Hebrew Bible

In the present contribution I want to focus on one aspect of Biblical Hebrew studies, for which the computational approach has proven to be very helpful: the linguistic diversity in the Hebrew Bible. This diversity has led to a broad range of explanations — differences in time of origin, genre, dialect, the influence of neighboring languages, individual styles of the authors, or transmission history.

I had the privilege to do my Ph.D. research (1995–1999) under the supervision of Professor Takamitsu Muraoka, who introduced me to this field of studies. My Ph.D. dissertation dealt with the verbal system in the Hebrew text of Ben Sira.[19] In my dissertation I compared the language of Ben Sira (presumably written in de beginning of the 2nd cent. BC) with the various forms of Biblical

[18] Cf. above, section 1; for further details see: Dirk Roorda and Wido van Peursen, "The Hebrew Bible as Data: Text and Annotations", Peter Boot, et al., eds., *Advances in Digital Scholarly Editing. Papers presented at the DiXiT conferences in The Hague, Cologne, and Antwerp* (Leiden: Sidestone, 2017), 323-331.

[19] A revised and enlarged version of the dissertation appeared as W. Th. van Peursen, *The Verbal System in the Hebrew Text of Ben Sira*, Studies in Semitic Languages and Linguistics 41 (Leiden: Brill, 2004).

and post-Biblical Hebrew. I concluded that the language of Ben Sira can be located in the development of the Classical Hebrew language between Biblical Hebrew and post-Biblical Hebrew (Dead Sea Scrolls, Mishnah), but that it is not merely a continuation of Late Biblical Hebrew, because it also contains some features it shares with Standard Biblical Hebrew as well as some unique features.[20] Hence, the reduction of the Hebrew book of Ben Sira as an unsuccessful attempt to imitate Standard Biblical Hebrew, as well as a purely chronological positioning of Ben Sira as an intermediate stage between Late Biblical Hebrew and Qumran Hebrew (or Mishnaic Hebrew) are both oversimplifications that do no justice to the unique linguistic profile of this Hebrew composition.

This research, as well as the series of conferences on the Hebrew of the Dead Sea Scrolls and Ben Sira that was initiated by Professor Muraoka, taught me the importance of extra-biblical and post-biblical sources for the study of language diversity and development within the Hebrew Bible as well as the value of syntactic analysis for describing this variation, rather than merely focusing on lexical or morphological issues.

At the end of the 20th century, when I completed my Ph.D. dissertation, the prevalent view was still that diachronic development was the main source of the linguistic variation in the Hebrew Bible, and that we could clearly distinguish between Archaic Biblical Hebrew, Early Biblical Hebrew (EBH), and Late Biblical Hebrew (LBH), to which Transitional Biblical Hebrew as an intermediate stage between Early and Late was added. It is the view that can be found in the reference works by E. Y. Kutscher (1982)[21] and A. Sáenz-Badillos (1993)[22] and that has been elaborated upon in more details by scholars such as A. Hurvitz.[23] However, since then, this view has been challenged by a number of scholars who questioned the methodological basis of the traditional view and advocated alternative approaches, such as the influential publication by Ian

[20] See W. T. van Peursen, "Chapter 6: Ben Sira", W. R. Garr and S. E. Fassberg, eds., *A Handbook of Biblical Hebrew*, 2 vols. (Winona Lake: Eisenbrauns, 2016), 1. 69-82, 2. 43-47.

[21] E. Y. Kutscher, *A History of the Hebrew Language* (Jerusalem and Leiden: Magness/Hebrew University and Brill, 1982).

[22] Angel Sáenz-Badillos, *A History of the Hebrew language* (Cambridge: Cambridge University Press, 1993).

[23] From his many publications on this topic see, e.g., his important study בין לשון לשון [*The Transition Period in Biblical Hebrew*] (Jerusalem: Bialik, 1972).

Young, Robert Rezetko and Martin Ehrensvärd (2008) on the linguistic dating of biblical texts.[24)]

4. A computational approach: the SynVar project

In 2013 Janet Dyk and the present author started a research project at the ETCBC on syntactic diversity in the Hebrew Bible.[25)] In this project, entitled "Does Syntactic Variation Reflect Language Change? Tracing Syntactic Diversity in Biblical Hebrew Texts",[26)] we purposefully refrained from taking an a *priori* stance in the debate described above (section 3) but started with the data itself: linguistic phenomena and their distribution throughout the entire Hebrew Bible and some extra-biblical sources (inscriptions, Dead Sea Scrolls, Mishnaic sources). Our intention was to use this distributional analysis to evaluate the broad range of explanations for the language variation in the Hebrew Bible.

The database allowed us to cover the complete Hebrew Bible, rather than one book or a small selection of books, as was often done in the past.[27)] The syntactic annotations of the ETCBC database allowed us to focus on syntax. Since syntactic structures are formed less consciously than words and phrases, they are less apt to imitation and manipulation, for example to create an archaizing style. Hence, to see what kind of shifts take place in a language and to decide whether or not we can discern chronological development, syntax is more trustworthy than lexical items, which may be the fruits of the conscious use of archaisms or purposeful imitations of a standard form of a language.

Three project constituents dealt with syntactic phenomena at the levels of

[24)] Ian Young, Robert Rezetko, and Martin Ehrensvärd, *Linguistic Dating of Biblical Texts*, 2 vols. (London: Equinox, 2008).

[25)] This project was made possible by the generous support of the Netherlands Organization for Scientific Research (NWO).

[26)] In the following we will use the abbreviation "SynVar".

[27)] E.g., on Chronicles: Robert Polzin, *Late Biblical Hebrew: Toward an Historical Typology of Biblical Hebrew Prose*, Harvard Semitic Monographs 12 (Missoula: Scholars, 1976); on Ezekiel: Mark F. Rooker, *Biblical Hebrew in Transition: The Language of the Book of Ezekiel*, JSOTSup 90 (Sheffield: JSOT Press, 1990); on Qohelet: Bo Isaksson, *Studies in the Language of Qoheleth with Special Emphasis on the Verbal System*, Studia Semitica Upsaliensia, (Uppsala: Almqvist and Wiksell international, 1987); on Esther: Ron Bergey, "Late Linguistic Features in Esther", *Jewish Quarterly Review* 75 (1984), 66-78.

phrases, clauses, and text hierarchical structures. The first two levels were studied in the Ph.D. projects of Marianne Kaajan and Martijn Naaijer. The third is in a postdoctoral project by Dirk Bakker.

The distribution of linguistic phenomena was analyzed according to various parameters. A first step was to investigate the distribution over the books or sections of the Bible. This could yield important insights, even though it would not necessarily lead to a compelling explanation. Differences between the alleged EBH and LBH corpora, for example, could point to the traditional labels of these corpora as 'early' and 'late', but they could also suggest different ways in which these corpora have been transmitted; if a phenomenon occurs most frequently in the Pentateuch, it might be that the careful transmission of the Torah is responsible for it; if we would see differences between narrative books like Samuel, legal instructions as those in Leviticus and the poetry of the Psalms, this may also point to different usages of the language, dependent on the genre.

However, when we come to genre, the parameter of text corpora (books or collections of chapters) is not enough and a refinement is needed. It is true that various types of communication show different usages of the language and that, for example, the language use of a narrative is different from that of a legal text or a sapiential instruction. However, for a linguistic analysis, a classification based on genre is not enough. The genre may suggest a certain text type (e.g. the patriarchal stories are "narratives"), but within a text, various text types may occur. In a story, for example, the characters may use discursive text in quoted direct speech. Sometimes the narrator addresses the listener or reader directly and switches from narrative speech to discursive speech. This happens, for example, when the narrator comments after the enigmatic episode of Jacob at Penuel: "Therefore to this day the Israelites do not eat the thigh muscle that is on the hip socket, because he struck Jacob on the hip socket at the thigh muscle" (Gen 32:32[33]). Here the narrator switches from the narrative register to addressing the reader/listener directly. For this reason we consider text type as a feature of a clause, which is assigned on the basis of syntax, rather than a feature of a larger literary unit, based on literary considerations. Within a literary unit all kinds of text type shifts can occur, which would be obscured if we define a text type as a feature of a verse, a paragraph or a larger literary unit.

The text types we distinguish are Narrative (N), Quotation (Q) and Discursive (D).

This categorization takes into account the difference between the narrative text type (N), the kind of communication that emerges when characters begin to speak within the narrative (Q), or the switch in communication that happens when the narrator addresses the reader/listener, as in the case of Genesis 31:13 mentioned above (D). For the statistical analysis of the distribution of linguistic phenomena, the categories N and Q appeared to be the most important.

In the analytical procedures for data enrichment in the ETCBC database, text types are automatically assigned on the basis of syntactic features. The deductions of the text types from syntactic features include rules such as: "If a clause contains a *wayyiqtol*, the text type is N." "If a clause contains a *yiqtol* and is preceded by a clause with text type N, the text type is ND."[28] "If a clause contains an imperative, a 1st or 2nd person inflection of pronoun or vocative, the text type is Q. If this occurs within a narrative context the text type is NQ."

In addition to the parameters of text corpora and text type, which mostly relate to language variation due to differences in language usage or language development, there is another parameter that is highly important from a linguistic perspective, namely the distinction between main and subordinate clauses. It is generally acknowledged that main clauses and subordinate clauses differ in their syntactic behaviour. Main clauses show more syntactic variation or 'liberty' than subordinate clauses. In the linguistic study of this phenomenon, a milestone is John Robert Ross' 1973 contribution. Ross coined the term "Penthouse Principle", which says that "More goes on upstairs than downstairs". "Upstairs" means here "within matrix clauses" (which in a tree representation usually occupy a higher node in the structures) and "downstairs" means "within subordinate clauses". Accordingly, when we study language variation, it is important to be aware that alternative expressions (e.g. the orders Verb-Subject and Subject-Verb) can be related to particular environments. Language change may start within one context (e.g. main clauses) and from there spread to other environments (e.g. subordinate clauses)

[28] The values for text type are constructed in a cumulative manner. Thus when within a narrative (N) a discursive text occurs, this is marked with ND; when within a narrative direct speech occurs, this is marked with NQ. When within a quotation another quotation occurs, this is marked NQQ. Hence the value of text type is a string of one or more characters, e.g. N, NQ, NQQ, ND etc.

5. Preliminary results

In the SynVar project, the distribution of linguistic phenomena was investigated according to the parameters described in the preceding section: text corpus, text type, and syntactic environment. Although the final deliverables of the project are still in preparation, already in this stage the research has yielded interesting insights and clearly indicated tendencies that can be observed at all three syntactic levels studied, those of phrases, clauses and text structures.

Perhaps the most important discovery of the research project is the interaction of the factors that are responsible for the distribution of linguistic phenomena.[29] That various factors play a role, such as genre or the difference between narrative sections and direct speech is commonly acknowledged. However, current research often focuses on such factors in isolation, focusing for example, on the linguistic features of Biblical poetry or the differences between the alleged EBH and LBH corpora. Our computational analysis demonstrated the interaction of these factors.

Thus, Marianne Kaajan, in her phrase level analysis, showed that there is a clear difference regarding the complexity of phrases between the alleged EBH and LBH corpora, but this difference appears only in the narrative text type (label: N), not in the direct speech sections (label: Q). She found some support for Polak's suggestion that early texts resemble spoken language, presumably due to the oral transmission of these texts, whereas in later text the narrative text type shows a more elaborate style, presumably reflecting chancery scribal practices.[30]

In her research of Biblical Hebrew phrase structure, Kaajan also developed new insights regarding the applicability of the notion of (non-)configurationality to Biblical Hebrew in relation to various kinds of discontinuous phrases. She also investigated the complexity of defining Biblical Hebrew 'main' and 'subordinate' clauses. In biblical scholarship the terms 'main' and 'subordinate' were often defined intuitively, with categories borrowed from English or other

[29] For our presentation of the preliminary results below we have drawn from the forthcoming Ph.D dissertations (Kaajan and Naaijer) and monograph (Bakker).

[30] Frank H. Polak, "The Oral and the Written: Syntax, Stylistics and the Development of Biblical Prose Narrative", *Journal of the Ancient Near Eastern Society* 26 (1998), 59–105.

familiar modern languages or from the grammatical tradition of Greek and Latin, but Kaajan developed a model that better fits the situation in Hebrew (and presumably other Semitic languages).

Martijn Naaijer, in his clause level analysis, demonstrated that for the realization of the copula in the EBH corpus the narrative text type (N) and the direct speech sections (Q) differ considerably, and that the latter (Quotation in EBH) shows similarities with the LBH corpus. In the LBH corpus the difference between Narrative and Quotation is much smaller. With respect to the use of the verb *haya* ‘to be’, the narrative text type shows similarities with main clauses.

Dirk Bakker, in his text level analysis, investigated the distribution and complexity of syntax trees. He found that in the alleged EBH corpus the verb form *wayyiqtol* on the average takes larger tree structures than *qatal*, but that this difference decreases in LBH, partly because of the decline of the *wayyiqtol* form. He concluded that many of the phenomena he observed can best be explained diachronically, but that in such an explanation the distinction between “innovative” main clauses and “conservative” subordinate clauses should be taken into account. Changes in the average size of tree structures take place in main clauses, and only later, or not at all, in subordinate clauses.

These observations support, but also modify the diachronic approach to the linguistic diversity in the Hebrew Bible. There are indeed observable difference between the alleged EBH and LBH corpora, but the status of these two collections of books differs in that the books in the EBH corpus shows much more linguistic homogeneity, whereas the books in the LBH corpus shows more internal diversity, while as a whole characterized by deviations from the EBH corpus.[31)]

The interaction of the distribution of linguistic phenomena over EBH and LBH with other parameters, such as text type (Narrative or Quotation) and syntactic context (main or subordinate clauses) suggests that linguistic innovations manifest themselves first in one particular context and from there may move to other contexts. From a linguistic perspective this agrees with the Penthouse Principle (see above).

[31)] This is one of the observations by Martijn Naaijer in his Ph.D. dissertation. See also Etienne P. van de Bijl, et al., “A Probabilistic Approach to Linguistic Variation and Change in Biblical Hebrew”, Victor de Boer, et al., eds., *Proceedings of the Network Institute Academy Assistants program 2017/2018* (e-book, 2019), accessed 22 April 2019 from http://doi.org/10.5281/zenodo.2546802.

Regarding the interaction between text type (N or Q) and presumed date of origin (EBH or LBH) there are contradicting tendencies at the various linguistic levels. On clause level (Naaijer) and text level (Bakker) one might hypothesize that direct speech sections in early narrative texts reflect somehow the spoken language and that in later texts the standards of written texts were increasingly influenced by spoken forms of the language. At phrase level (Kaajan), however, we see the opposite phenomenon: in the narratives in the early texts, both text types N and Q seem to reflect oral language, while in later texts only Q sections keep reflecting oral language and N sections develop towards a more complex, presumably scribal, style.

6. Conclusion

Applying digital research methods to the Hebrew Bible supports a systematic investigation of linguistic phenomena, their distribution, and their interpretation and translation. At the basis of the development of the database, there was the underlying concern for the text as it stands and the patterns that can be discovered in it, which should not be overruled by thematic or theological considerations. In that sense it could even be described as a modern application of the *sola scriptura* principle.[32] The digital analysis of the Hebrew Bible has proven useful for the study of modern and ancient Bible translations as well as for the study of the Hebrew Bible as a linguistic corpus with its fascinating distribution of linguistic patterns and phenomena.

<Keywords>
Hebrew linguistics, Syntactic variation, Digital Humanities, Narrative and direct speech.

(투고 일자: 2019년 1월 31일, 심사 일자: 2019년 2월 25일, 게재 확정 일자: 2019년 4월 16일)

[32] Cf. Kim Kyoungsik, “ETCBC Data for the Libre Bible Software”.

<References>

Bergey, Ron, “Late Linguistic Features in Esther”, *Jewish Quarterly Review* 75 (1984), 66–78.

Bijl, Etienne P. van de, et al., “A Probabilistic Approach to Linguistic Variation and Change in Biblical Hebrew”, Victor de Boer, et al., eds., *Proceedings of the Network Institute Academy Assistants program 2017/2018* (e-book, 2019), accessed 22 April 2019 from http://doi.org/10.5281/zenodo.2546802.

Doedens, Crist-Jan, *Text Databases: One Database Model and Several Retrieval Languages*, Language and Computers 14, Amsterdam and Atlanta, GA: Rodopi, 1994.

Dyk, J. W., “Deportation or Forgiveness in Hosea 1:6? Verb Valence Patterns and Translation Proposals”, *The Bible Translator* 65:3 (2014), 235–279.

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

Hurvitz, Avi, בין לשון לשון [*The Transition Period in Biblical Hebrew*], Jerusalem: Bialik, 1972.

Isaksson, Bo, *Studies in the Language of Qoheleth: with Special Emphasis on the Verbal System*, Studia Semitica Upsaliensia 10, Uppsala: Almqvist and Wiksell international, 1987.

Keulen, P. S. F. van and Peursen, W. Th. van, eds., *Corpus Linguistics and Textual History: A Computer-Assisted Interdisciplinary Approach to the Peshitta*, Studia Semitica Neerlandica 48, Assen: Van Gorcum, 2006.

Kim, Kyoungsik, “ETCBC Data for the Libre Bible Software”, http://etcbc.nl/uncategorized/etcbc-data-for-the-libre-bible-software/ (22 April 2019).

Kingham, Cody, “Data Creation”, http://www.etcbc.nl/datacreation/ (22 April 2019).

Kutscher, E. Y., *A History of the Hebrew Language*, Jerusalem and Leiden: Magness/Hebrew University and Brill, 1982.

Oosting, Reinoud, “Computer-Assisted Analysis of Old Testament Texts: The Contribution of the WIVU to Old Testament Scholarship”, Klaas Spronk, ed., *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 Oudtestamentisch Werkgezelschap*, Leiden: Brill, 2016, 192–209.

Petersen, Ulrik, “Emdros – A Text Database Engine for Analyzed Or Annotated Text”, Geneva: *Proceedings of the COLING Conference*, 2004. Available online: https://emdros.org/petersen-emdros-COLING-2004.pdf.

Peursen, Wido van, "Chapter 6: Ben Sira", W. R. Garr and S. E. Fassberg, eds., *A Handbook of Biblical Hebrew*, 2 vols., Winona Lake: Eisenbrauns, 2016.

Peursen, Wido van, "'This is What Was Spoken by the Prophet Joel'. The Latter Rain in Joel's Prophecies and in Dutch Pentecostalism", M. Klaver, S. Paas, and E. van Staalduine-Sulman, eds., *Evangelicals and Sources of Authority*, Amsterdam Studies in Theology and Religion 6, Amsterdam: VU University Press, 2016, 271-285.

Peursen, W. Th. van, *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, 2007.

Peursen, W. Th. van, "Progress Report: Three Leiden Projects on the Syriac Text of Ben Sira", R. Egger-Wenzel, ed., *Ben Sira's God. Proceedings of the Second International Ben Sira Conference, Durham, Ushaw College, 2001*, Beihefte zur Zeitschrift für die Alttestamentliche Wissenschaft 321, Berlin: De Gruyter, 2002, 361-370.

Peursen, W. Th. van, *The Verbal System in the Hebrew Text of Ben Sira*, Studies in Semitic Languages and Linguistics 41, Leiden: Brill, 2004.

Polak, Frank H., "The Oral and the Written: Syntax, Stylistics and the Development of Biblical Prose Narrative", *Journal of the Ancient Near Eastern Society* 26 (1998), 59-105.

Polzin, Robert, *Late Biblical Hebrew: Toward an Historical Typology of Biblical Hebrew Prose*, Harvard Semitic Monographs 12, Missoula: Scholars, 1976.

Rooker, Mark F., *Biblical Hebrew in Transition: The Language of the Book of Ezekiel*, JSOTSup 90, Sheffield: JSOT Press, 1990.

Roorda, Dirk and Peursen, Wido van, "The Hebrew Bible as Data: Text and Annotations", Peter Boot, et al., eds., *Advances in Digital Scholarly Editing: Papers presented at the DiXiT conferences in The Hague, Cologne, and Antwerp*, Leiden: Sidestone, 2017, 323-331.

Sáenz-Badillos, Angel, *A History of the Hebrew language*, Cambridge: Cambridge University Press, 1993.

Talstra, Eep, "Text, Tradition, Theology: The Example of the Book of Joel", E. Van der Borght and P. van Geest, eds., *Strangers and Pilgrims on Earth: Essays in Honour of Abraham van de Beek*, Leiden: Brill, 2011, 309-327.

Young, Ian, Rezetko, Robert, and Ehrensvärd, Martin, *Linguistic Dating of Biblical Texts*, 2 vols., London: Equinox, 2008.

<Abstract>

A Computational Approach to Syntactic Diversity in the Hebrew Bible

Wido van Peursen
(Vrije Universiteit Amsterdam)

For more than four decades, the Eep Talstra Centre for Bible and Computer (ETCBC) has been building a richly-annotated linguistic database of the Hebrew Bible. This contribution describes the processes of data creation of this database and its underlying methodological principles. These principles which can be labeled as “bottom-up” and “form-to-function” stem from a deep concern to do justice to the biblical text itself and to prevent it from being overruled by thematic or theological considerations.

The database facilitates the application of computational linguistics and digital humanities to the Hebrew Bible, and supports biblical exegesis, Bible translation as well as the study of the Bible as a language corpus. In recent years, the ETCBC database has been transformed to an open tool, which can be consulted online and be downloaded as a package for anyone who wants to use it for more advanced computational analysis of the Hebrew Bible.

A research project on syntactic variation in the Hebrew Bible demonstrated the interaction of presumed data of origin (e.g., early versus late texts), genre (e.g., prose or poetry), text type (e.g., narrative and direct speech), and syntactic environment (e.g., main versus subordinate clauses). Regarding the realization of the copula “to be” for example, it can be observed that the narrative text type and the direct speech sections differ considerably in the alleged early texts of the Bible and that the direct speech in the early corpus shows similarities with the Late Biblical Hebrew corpus.

Regarding the complexity of tree structures, it can be observed that changes in the average size of tree structures take place in main clauses, and only later, or not at all, in subordinate clauses. This agrees with a well-known principle in linguistics, the so-called Penthouse Principle, that accounts for the distinction between “innovative” main clauses and “conservative” subordinate clauses.

Such distribution patterns which can only be discovered with a computational

full corpus analysis are helpful to get a better understanding of diachronic language development of Classical Hebrew in the intersection of oral and written text transmission.