Article

New Directions in the Computational Analysis of Biblical Hebrew Grammar

Cynthia L. Miller-Naudé
University of the Free State,
South Africa
millercl@ufs.ac.za

Jacobus A. Naudé
University of the Free State,
South Africa

Abstract

The concern of the paper is to highlight how computational analysis of Biblical Hebrew grammar can now be done in very sophisticated ways and with insightful results for exegesis. Three databases, namely, the Eep Talstra Centre for Bible and Computer (ETCBC) Database, the Accordance Hebrew Syntactic Database, and the Andersen-Forbes Syntactic Database, are compared in terms of their relation to linguistic theory (or, theories), the nature and spectrum of retrieved data, and the representation of synchronic and diachronic linguistic variation. Interaction between different contexts, including the African context, are promoted namely between linguists working on Biblical Hebrew and exegetes working on the Hebrew Bible by illustrating how exegesis and language are intimately connected, as well as among geographical contexts by comparing a European database (ETCBC), a North American database (Accordance) and a Southern hemisphere database (Andersen-Forbes).

Keywords: corpus linguistics; electronic text analysis; linguistic theory; linguistic variation; diachrony

Introduction

The exegesis of the Hebrew Bible ultimately depends upon understanding its language. The linguistic study of Biblical Hebrew grammar has been revolutionised in recent years through the development of computerised databases. Initially, the databases were capable of retrieving only lexical and morphological data, but in recent years databases have been developed for the retrieval and analysis of syntactic data and even data on the

Journal for Semitics
https://upjournals.co.za/index.php/JSEM
Volume 27 | Issue 1 | 2018 | #4628 | 17 pages

https://doi.org/10.25159/1013-8471/4628
ISSN 1013-8471 (Print)

discourse level from the Hebrew Bible. Computational analysis of the Hebrew Bible can now be done in very sophisticated ways and with insightful results for exegesis.

In the light of these developments the authors of this article were invited to organise a panel with the theme "New Directions in the Computational Analysis of Biblical Hebrew Grammar" at the 2016 meeting of the International Organization for the Study of the Old Testament (IOSOT) which was held at the University of Stellenbosch.[1] Representatives from three database projects were invited to demonstrate their approaches to the computational analysis of Biblical Hebrew grammar: Wido Van Peursen (Free University of Amsterdam) of the Eep Talstra Centre for Bible and Computer (ETCBC) Database, Dean Forbes (University of the Free State) of the Andersen-Forbes Syntactic Database, and John A. Cook (Asbury Theological Seminary) of the Accordance Hebrew Syntactic Database. These databases represent differing approaches to computational analysis of the Hebrew Bible, as well as geographical diversity (Europe, Southern Hemisphere, and North America).

The retrieval of syntactic data, however, opens up new theoretical and practical questions. Among them are the following: (1) which linguistic theory or theories are used to tag the data? (Or, conversely, which linguistic theories are accommodated within the database?) What are the implications of the choice of linguistic theory for the retrieval of data? (2) What kinds of grammatical data can be retrieved? What kinds of grammatical data remain elusive? (3) To what extent can features that exhibit linguistic variation—both synchronic variation (or, "style") and diachronic variation—be retrieved?

Representatives from the three database projects address these questions in three articles following this article, namely (in order of oldest to youngest database): A. Dean Forbes and Francis I. Andersen (University of the Free State) of the Andersen-Forbes Syntactic Database, Cody Kingham and Wido Van Peursen (Free University of Amsterdam) of the ETCBC Database, and Robert D. Holmstedt (University of Toronto) and John A. Cook (Asbury Theological Seminary) of the Accordance Hebrew Syntactic Database. In this article, we contextualise the articles on the specific databases that follow. In the first section, we briefly introduce the issue of syntactic databases within the context of corpus linguistics. We then provide an overall comparison of the databases with respect to their approach to linguistic theory, the kinds of grammatical data that they can retrieve, and the extent to which they can retrieve synchronic (or "stylistic") variation and diachronic variation. Finally, we envision possible future enhancements to the databases.

[1] The panel was linked on the IOSOT programme to the plenary presentation of Wido Van Peursen on the computational analysis of biblical Hebrew poetry. The published version of the paper is Van Peursen (2017).

Syntactic Databases, Corpus Linguistics and Electronic Text Analysis

Linguistics as a discipline is the science of language which studies linguistic reality. It takes as its object universal aspects of language structure and function (to create a general linguistic theory), as well as the description and comparison of individual languages (to create a grammar of a particular individual language). Linguists use their knowledge of a specific language to enhance their understanding of language as such; conversely, they apply that general understanding to the study of a specific language, the grammar of a specific language, which describes linguistic competence and explains the products of a specific language. This is done by using well-articulated linguistic methods and applying general linguistic theories to particular constructions or corpora of that specific language (Botha 1981, 432-439).

Although the goal of linguistic inquiry is the description of a language system, the linguist has no direct access to it (Miller 2004, 282). It is only by instantiations or products of language use that linguists are able to discern the abstract structures of language or the competence of a native speaker. Native speakers have phonological, morphological, syntactic, semantic and pragmatic competence and this competence is reflected in their intuitions about these aspects of their native language. Linguists working in living languages use native speakers in three ways: for eliciting specific linguistic expressions for analysis, for judgments about grammaticality of linguistic expressions and to provide metalinguistic intuitions about the structure or function of the language (Miller 2004, 291). Another approach for collecting the data with which to discern the language system involves the analysis of collections of texts.

Advances in information technology and software development have resulted in the use of electronic resources to complement manual approaches to the analysis of language and literature.[2] This means that data can be manipulated in ways that are simply not possible otherwise. Many of the techniques used in the electronic analysis of texts

[2] The information technology revolution has changed the composition of work as personal computers, cell phones, internet and their social-media offshoots have spread. The impact is strengthened by the merger of globalisation and the information technology revolution that coincided with the transition from the twentieth to the twenty-first century (Friedman 2005). After 2005 there was the move to universal connectivity to the internet via cell phone and smartphone, in addition to the personal computer. This connectivity is being supported by a vast new array of software applications stored on huge interlinked server farms known collectively as “the cloud” (Friedman and Mandelbaum 2011, 59–65). Any individual user’s device is now turned into an information-creation or information-consumption powerhouse in a hyper-connected world, which “constitutes the most profound inflection point for communication, innovation, and commerce since the Gutenberg printing press” (Friedman and Mandelbaum 2011, 64). Though the effects of the printing press took hundreds of years to percolate through society, hyper connectivity happened in more or less a decade, which demands more challenges in terms of adaptation (Friedman and Mandelbaum 2011, 64).

originate from manual procedures of text analysis which were used before the more recent advent of computer technology. An example is the manual concordance extraction of selected items in the Bible.[3] The first electronic text analysis tools were designed in the 1950s and initially only produced paper concordances.[4]

The area of enquiry of computer-aided language research, which is referred to as *corpus linguistics,* involves the development of principled collections of electronic texts or corpora based on large sets of naturally occurring language data as well as the systematic exploration of recurring patterns in the use of language by corpus tools with the aim of gaining a better understanding of language in use.[5] Biber, Conrad and Reppen (1998, 246–250) set three essential design criteria which a linguistic corpus of a living language must satisfy (see also Biber 1990). First, the corpus must be extensive so that the relevant linguistic features are present within the texts to be analysed. Second, the corpus must be representative, both with respect to the varieties of language contained in it and with respect to the kinds of texts that are included. Third, the corpus must be varied—an extensive, monolithic corpus is not as valuable as a smaller, diverse one. One can carry out more sophisticated linguistic investigations with what are known as “marked up” and/or “annotated” texts. Mark-up involves inserting tags to make explicit the appearance and structure of a text, while annotation involves inserting tags in order to make explicit the linguistic features (part of speech, syntactic structure) of a text (Bowker and Pearson 2002, 75–91, McEnery and Wilson 2001, 131, Meyer 2002, 81–99). Taggers have been designed to insert part of speech tags automatically. A tagged corpus is useful to investigate classes of words (for example adjectives) rather than individual words in a corpus. There are two main approaches in corpus design: the corpus-based, where the researcher uses the data provided by the corpus to test previously established hypotheses and corpus-driven, where the researcher aims to derive linguistic categories systematically from the recurrent patterns and the frequency distributions that emerge from language in context (Teubert and Čermáková 2007, 137). The results of research in corpus linguistics are applied in various areas of language

[3] Teubert and Čermáková (2007, 9–11, 137) view the methodology of corpus linguistics as similar to that of philologists in that the grammar rules developed by philologists were specific rules to make sense of texts of a particular language.

[4] See Hughes (1988, 343–383) for the development of electronic Bible concordance programs.

[5] Other traditions and methodologies of computer-aided language research include Natural Language Processing (NLP), which studies how computers can be made to process and interpret naturally occurring text for the design of applications for example spell checkers and machine translation software, and Humanities Computing, which investigates how technology can be used for research into humanities subjects by documenting textual interpretations through computer-based analysis and annotation. These applications produce outputs that have relevance outside of linguistics. The research goal is the successful development of an application rather than the comprehensive description of language in use. For computer-assisted language learning, see Hughes (1988, 385–428).

study.[6] These include, amongst others, language description and analysis,[7] comparisons between texts and text collections as part of authorship studies,[8] the study of how ideology is encoded in language,[9] and the exploration of corpus data for language teaching applications.[10]

The core research activity in corpus linguistics is the extraction of language patterns through the analysis of sorted instances of particular lexical items and phrases. This activity provides a different look at language which is claimed to be representative of actual language use (Biber, Conrad and Reppen 1998, 1). Linguists had become dissatisfied with the insufficient descriptions for the various languages, where grammar rules in these descriptions are violated in texts with the result that they began to utilise corpora to compile more adequate descriptions.

A corpus linguist organises language/textual data through the generation of frequency information, which represents individualwords or phrases in a concordance format, followed by the analysis of concordance lines (McEnery, Xiao and Tono 2006). Basic information about a text or collection of texts can be retrieved by the concordance tool, including sentence length, word length, number of paragraphs, ratio between the number of running words in a text and the number of different words in a text (type-

[6] Prominent work is done by Sinclair (1991, 2003, 2004). The standard English grammar of Quirk et al. (1985) is based on manually collected language data and was the first large-scale project to collect language data for empirical grammatical research. Since the mid-1980s, the data are computerised in Quirk and Greenbaum's subsequent project, known as the International Corpus of English (ICE) (http://www.ucl.ac.uk/english-usage/ice/). The second data-oriented project in the 1960s was the Brown Corpus, named after Brown University in Providence, Rhodes Island (http://icame.unib.no/brown/bcm.html) (Teubert and Čermáková 2007, 50–51).

[7] Corpus linguistics facilitates the study of both synchronic and diachronic variation by making it possible to trace language changes even over short periods of time. Biber, Conrad and Reppen (1998, 21–230) discuss the impact on the investigation of language features (lexicography, grammar, lexico-grammar, discourse) and on the investigation of characteristics of varieties (register variation, language acquisition and development, as well as diachronic and stylistic variation). See also McEnery and Wilson (2001, 103–132) for the use of corpora in language studies.

[8] Forensic linguistics combines corpus linguistic methods with statistics to uncover authorship and plagiarism. In corpus translation studies parallel or comparable corpora (also known as translation corpora), which contain equivalent and usually aligned texts in two or more languages, are sometimes utilised (Olohan 2004, 24–34). See Lombard and Naudé (2009), Marais and Naudé (2007), Naudé (2004), Naudé (2008), Snyman, Ehlers and Naudé (2007) for various applications of corpora in translation studies and translation technology.

[9] Electronic text analysis has been used to study gender-related language and swearing in discourse (see also Adolphs 2006, 80–96).

[10] A key advantage of using corpora in language teaching is that they provide actual evidence of language use in different discourse contexts for the language learner and not the made-up language usage in traditional teaching materials that are based on intuition (see also Adolphs 2006, 97–116).

token ratio) (Adolphs 2006, 80–96; Oakes 1998). Word lists can be generated in different rank orders including alphabetical, frequency, part of speech and lemma (Adolphs 2006, 80–96). This information provides an initial picture of the consistency of texts and in this regard **keywords** play a central role, which have three references in corpus linguistics. First, it is the search word/search term in a concordance study. Secondly, it refers to those items that occur with a significantly higher or lower frequency in a text when compared with the larger reference corpus. Thirdly, it refers to lexical items or phrases that have a particular relevance to a research topic. The Key Word in Context (KWIC) concordance has become a standard way of presenting instances of individual lexical items and phrases in a given text or text collection, where the search word or keyword appears in the middle of the line with its co-text on either side. This representation facilitates the analysis of lexical and grammatical patterns in the immediate environment of the search term.

Electronic text analysis is any type of analysis that draws on digitised texts including the process of adding analytical and contextual information to such texts. It is not confined to the methods developed and used within the area of corpus linguistics. Using electronic text analysis to study naturally occurring discourse is a replicable process; the analysis can be verified by other researchers. It also allows the language data to be manipulated in various ways to suit particular research purposes. Electronic text analysis helps to identify patterns that do not tend to be open to intuitive inspection, for example information about word frequency and co-occurrence of particular words. Electronic text analysis can be used at different stages in the analytical process. For example, frequency lists may be used in a quantitative way which may lead to a subsequent qualitative exploration or they may be used after an initial qualitative exploration. Electronic text analysis is often used as a complementary approach in research.

Electronic text analysis also has limits when it is used in isolation. Electronic text analysis cannot easily handle representations that are not textual, such as resources that include visual and audio elements. Similarly, intonation and body language in spoken discourse cannot easily be analysed with methods of electronic text analysis. Furthermore, the occurrence of a particular word or phrase in a corpus is not necessarily an indication of its frequency in the language in general. Corpus research can only produce results that reflect the particular corpus that is being used for a study.

Comparison of the Biblical Hebrew Databases

Relation to Linguistic Theory

It is impossible to create a syntactic database without some approach to or understanding of linguistic theory. In determining what to tag and what labels to use, an insight into syntactic theory is critical. Linguistic theory provides a lens through which to isolate, identify, describe and catalogue syntactic structure. Various linguistic theories, however, provide significantly different ways of identifying and describing syntactic structures. Linguistic theories are also constantly changing and evolving. It is impossible for a database to change as quickly as linguistic theories do, and it is not desirable. The relationship of the databases to linguistic theory is thus a very interesting question. Each of the databases has answered the question differently, in large part based upon the linguistic preferences of their creators, and yet they show similarities in some important respects.

The Andersen-Forbes syntactic database is the oldest database, begun in 1970. Although Chomsky's generative theory was (and remains) the dominant linguistic theory, Andersen knew Kenneth Pike and was drawn to structuralism as a result of his tagmemic theory (e.g., Pike 1967; Pike and Pike 1977). Andersen drew upon aspects of tagmemics for his analyses of the verbless clause (Andersen 1970) and the sentence in Biblical Hebrew (Andersen 1974). Forbes knew linguists who worked within another linguistic theory, Generalised Phrase Structure Grammar (GPSG) (Gazdar, Klein, Pullum and Sag 1985, Gazdar 1988), which was also a reaction against generative linguistics, and drew upon that theory. The linguistic approach that was eventually developed and used by Andersen and Forbes is non-generative and structuralist. Although their approach is eclectic and even idiosyncratic in a number of respects, it is informed by a wide spectrum of linguistic theories. Their volume *Biblical Hebrew Grammar Visualized* (Andersen and Forbes 2012) provides a detailed description of the database, of the linguistic theories that underlie it, of the metalanguage that they employ, and of various linguistic decisions that they made in developing the database and the reasons for them.

The Andersen-Forbes approach to linguistic structure is based on surface structure. It is constituent-based and monostratal. It is a flat structure in the sense that rather than employ hierarchical structures such as VP [Verb Phrase], a constituent involving the verb and its complements (objects) and adjuncts, all of the constituents in a clause/sentence are on the same level, as nodes of a single tree. However, the structure that they represent does allow for certain kinds of complex structures. Discontinuous structures, ones in which a constituent occurs in more than one piece in the structure (e.g., a discontinuous appositional structure), are represented. They also allow multidominance of nodes which allows a constituent to operate across clauses, as for example, with gapping (ellipsis). They focus on data rather than theory. One illustration of the emphasis on theory is the way in which they tag participles. They identify four kinds of participles based upon their syntactic function: (1) pure noun participle (participles which display only nominal characteristics); (2) pure verb participle (participles which display only verbal characteristics); (3) noun-verb participle (participles which display nominal characteristics with a previously occurring segment, but verbal characteristics with what follows); (4) noun [verb/noun] participle (participles which are in mixed constructions, displaying both verbal and nominal characteristics with following segments) (see Andersen and Forbes 2012, 33–35). This kind of attention to the syntactic relationships between participles and their syntactic context resulted in new grammatical categories and provided new insights into Biblical Hebrew syntax.

The ETCBC database began in 1977. Like the Andersen-Forbes database, it is structuralist in its orientation to linguistic theory (especially Wolfgang Schneider and Harald Weinrich; see Talstra 1978, 1992; Schneider 2007), although in recent years cognitive linguistics has also played a role. By beginning with morphological data at the word level, the encoding of information identified morphological information and then moved from the "bottom-up" to phrases, sentences and ultimately discourse. A second guiding principle of the database is "form-to-function". This means that form is primary and function is assigned only after all of the data with the same form are collected and analysed. The database used automatic processing of rules to process grammatical units. *Distributional units* are those which can be identified through "formal pattern recognition"; by contrast, *functional units* are those which can only be identified with human interpretation. The traditional constituents of phrases, clauses and sentences are functional units. At the highest level of the bottom-up analysis, are clause relations that make up the discourse structure of the text.

The Accordance database is the newest database—it began ten years ago. One of the reasons for beginning a new database was the concern to have a database that was more explicitly connected to linguistic theory, especially generative linguistics, and focused on syntax. At the same time, the founders of the database wanted a tool that could be used by a wide variety of scholars rather than exclusively for generatively-oriented scholars. They therefore adopted the motto "data primary, theory wise".

The generative orientation of the database can be seen in the following decisions concerning analysis and identification of constituents. First, they follow the generative principle that every phrase constituent has a "head"; omission of the head results in a null constituent (e.g., an implicit subject with a finite verb or the covert copula in verbless clauses). Second, they distinguish complements of the verb from adjuncts on the basis of the valency of lexical verbs. Third, they use a hierarchical (as opposed to flat) structure, thus identifying verb phrases (or predicate phrases) even though verb phrases in Hebrew are often discontinuous since the verb is often in initial position and separated from its complement and adjunct(s) by the subject. It is important to note, however, that like the other two databases and in contrast with current generative linguistic theory, they do not require binary structures. Fourth, they decided not to represent movement of constituents (e.g., from topicalisation) in the database, but instead allow discontinuous constituents which are the result of movement. Discontinuous constituents are bound together, so to speak, by a system of cross-referencing. Cross-referencing also allows them to represent dislocation (*casus pendens*) constructions, as well as resumptive elements in relative clauses and ellipsis. Finally, apart from judgements about the valency of lexical verbs, which is semantically based, they do not encode other kinds of semantic or pragmatic information or discourse relationships. In this respect, the database is focused on syntax.

The creators of each of the databases have produced their databases in a manner that is informed by linguistic theory and none of them adheres slavishly to theoretical concerns. It is most interesting that none of them have opted for a binary approach to syntactic analysis, not even the Accordance database, which is centrally informed by generative linguistic theory. The linguistic choices made by the database creators are best evident in their approaches to the verb phrase as a constituent, to ellipsis, in the use (or non-use) of semantic or pragmatic labels, and in the extent to which they are interested in representing higher levels of syntax (discourse and participant tracking across sentences).

Nature and Spectrum of Retrieved Data

It is worth remembering that when the Andersen-Forbes and ETCBC databases began, computers were large pieces of machinery owned by universities and corporations, but not by individuals (see Forbes 2014 for a description of the early equipment used). Furthermore, both databases first had to devise means to represent the Hebrew text electronically and then to segment it and analyse it. Prior to this time, concordances of words had to be manually collected and printed; collections of syntactic structures for analysis and inclusion in grammar books could only be done by reading the text. The two databases segmented and encoded the morphological data differently in some respects (see Hughes 1988, 498–509). The ETCBC database follows a strictly morphological approach, segmenting all derivational and inflectional morphemes. Andersen-Forbes rather use the text “segment” as the basic unit: “A segment can be a word (“free morpheme”), a part of a word (“a bound morpheme”), or a sequence of words” (Andersen and Forbes 2012, 15). In many instances, a segment is a morpheme. The main distinction from a strictly morphemic analysis occurs in verbs, where the subject pronoun affixes are not segmented, and in some proper nouns (e.g., בֵּית־אֵל) and a few conjunctions (e.g., כִּי־אִם) which are “ligatured” so that they are handled as one segment (Andersen and Forbes 2012, 15–17).

All the databases are able to retrieve graphemic, phonological and morphological information. However, they differ in terms of their representation of the Masoretic accents (cantillation). The ETCBC database includes no cantillation information. Andersen-Forbes chose not to represent the cantillations, although they used the information from the accents to distinguish forms which are otherwise identical (e.g., absolute versus construct of segholate nouns) (see Andersen and Forbes 2012, 330–331). Only the Westminster Hebrew morphology database (Groves-Wheeler Hebrew Morphology v. 4.20), which forms the morphological underpinnings of the Accordance database, includes full representation of the Masoretic accents.

The three databases differ in the textual materials included within them. All the databases include the text of the Hebrew Bible, including the Aramaic portions. The Andersen-Forbes and ETCBC databases produced this information themselves and thus had to make decisions concerning which Hebrew text to represent, how to handle *Ketiv-Qere*, and whether to include Masoretic accents, whereas the Accordance database was layered on top of the Accordance Westminster Hebrew Bible information.

Beyond the Hebrew Bible, the Accordance database focuses on Hebrew and includes all ancient Hebrew texts, from epigraphic texts to the Dead Sea Scrolls. The ETCBC database includes, in addition to the Hebrew Bible, selected non-biblical texts: (1) epigraphic texts, both in Hebrew (Siloam Inscription, Kuntillet Adjud, Arad, Lachish, Mesad Hashavyahu, Ketef Hinnom amulet) and in other Northwest Semitic languages (Mesha Stele, Deir Alla); (2) selected non-biblical Qumran texts (the War Scroll [1QM], the Community Rule [1QS], 4Q246);[11] (3) selected Tannaitic texts (the Mishnaic tractate Avoth and the Parasha Shirata from Mekilta d-Rabbi Ishmael), selected Syriac texts (Peshitta of Kings, Judges, Ben Sira, the Prayer of Manasseh, Epistle of Baruch, and the Book of the Laws of the Countries); and (4) selected Targumim (Targum Jonathan on Judges). For research involving diachronic features of Hebrew, it is a desideratum for the databases to continue to expand their coverage to all pre-modern Hebrew texts. For research involving comparative Northwest Semitic philology, the ETCBC database provides the most expanded coverage beyond the Hebrew Bible, both in terms of languages and in terms of time frame, although it is not yet comprehensive.

The choices concerning linguistic theory and representation as described above in the previous section have implications for the kinds of data which can be retrieved. A few examples will be given here. First, the Andersen-Forbes database provides 28 semantic labels for common nouns, including “unknown/undecidable”.[12] These were initially incorporated to avoid subject/objection confusion, but can be useful in compiling syntactic constructions involving, for example, common nouns as “human” or “natural material” or “abstract quality”. The other two databases do not provide this information.

Second, the ETCBC Database uses a conventional set of 13 parts of speech based upon those in Köhler and Baumgartner (2001): verb, noun, proper noun, adjective, adverb, preposition, conjunction, personal pronoun, demonstrative pronoun, interrogative pronoun, interjection, negative particle, and interrogative particle. Andersen-Forbes developed a highly differentiated parts-of-speech system (2012, 20–42) with 76 parts-of-speech. By contrast, Accordance labels constituents with respect to their syntactic role in the clause—core constituents (e.g., subject, predicate [or verb], complement,

[11] Work in progress includes the following Qumran texts: Hodayoth (1QH^a), Habakkuk Pesher, and the Temple Scroll (11QT^a).

[12] Semantic classes for proper nouns (e.g., ethnic group, location) will be added to the database.

adjunct) and non-core constituents (vocative, exclamative/interjection, parenthesis, apposition, casus pendens).[13]

Third, the ETCBC database provides a wealth of information on discourse structure, which is built up from the clause structure. They indicate a limited number of clauses (based on type of verb/predicate and order of clause constituents). They indicate sentence boundaries. They indicate whether a text is narrative, discourse, or embedded and the relations between clauses in a tree hierarchy. They have experimental work on "paragraphs" in Biblical Hebrew and they have extensive information on participant tracking. The Andersen-Forbes database is working towards the incorporation of "text-types" into the data. They want a unified syntax-discourse transition in which clauses play a direct role over a disjoint one. They do, however, indicate "exchange structure" (the adjacency pairs of dialogue). The Accordance database does not encode discourse information. They do, however, indicate the quotation of direct speech as the embedded complement of the verb of speaking and their coindexation of participants allows for much information that is relevant to discourse to be retrieved.

Fourth, as mentioned above, only the Accordance database includes verb valency in its representation. Adding valency information is planned for the ETCBC database.

Fifth, the Accordance database allows for easy retrieval of constructions involving null constituents—ellipsis, verbless clauses (null copula), headless relatives. It is also possible to retrieve these constructions with the other databases, which do not directly encode null constituents.

Representation of Synchronic and Diachronic Variation

By linguistic variation, we refer to variation involving alternative, grammatically acceptable linguistic structures (see Miller-Naudé 2012). Synchronic variation (sometimes referred to as "style") is linguistic variation that relates either to the demographics of the speaker or to a particular register of language use. Diachronic variation is variation that is attested over time; it often grows out of synchronic variation.

Numerous kinds of linguistic variation are present in the Hebrew Bible at all levels of linguistic structure.[14] The databases differ in the kinds of variation that can easily be retrieved. Some kinds of variation relate to features of the Masoretic text. *Ketiv-Qere*

[13] See Miller-Naudé and Naudé (2017) for a discussion of linguistic approaches to grammatical categorisation (parts of speech) in Biblical Hebrew.

[14] Another kind of variation relates to alternative syntactic analyses of the data. The Andersen-Forbes database explicitly encodes ambiguous syntactic structures (e.g., Amos 1:2; see Andersen and Forbes 2012, 310), preferring to represent two possible analyses rather than choosing one.

variants can be retrieved from all the databases, as can orthographic variants. There is also variation caused by “obvious errors” in Codex Leningradensis; the Andersen-Forbes database corrects these (Andersen and Forbes 2012, 328–330). Morphological variation and lexical variation can also be easily retrieved from all the databases.

Syntactic variation is of many kinds; to the extent that it depends upon morphological or lexical features, it can be easily retrieved from all the databases. A few examples will suffice to illustrate. There is alternation in the kind of object (complement) that a transitive verb takes—bare noun phrase, noun phrase introduced with the definite object marker, objective suffix or prepositional phrase. The Accordance database explicitly encodes complements (objects) of verbs and can retrieve these data easily. Similarly, some lexical verbs exhibit alternation in the specific preposition used with them; see the discussion of the alternation of the prepositions אֱל and ל to introduce the indirect object after the verb אמר in Andersen and Forbes (2012, 344–345).

Variation involving language register or genre can be retrieved easily from the ETCBC database, which differentiates narrative and discourse, and from the Andersen-Forbes database which distinguishes narration, indirect speech, dialogue and exposition (Andersen and Forbes 2012, 313, 356–358). The Andersen-Forbes database also differentiates who is speaking in dialogic exchanges (e.g., human, deity). Another potential source of variation might be the sources underlying the biblical text; Andersen-Forbes tags Eissfeldt’s hexateuchal sources, thus allowing analysts to explore how those sources relate to syntactic variation (Andersen-Forbes 2012, 355–356).[15]

Diachronic change involves variation over time. One of the reasons for the creation of the Accordance database was to “allow deeper research into diachronic syntactic development” by being able to retrieve the syntactic structures of the “full scope of ancient Hebrew texts” (Holmstedt and Cook 2018). In other words, they wanted to be able to trace the trajectories of syntactic change from the earliest Hebrew texts to the Qumran texts. The ETCBC database extends the possible trajectory further by inclusion of some Tannaitic texts.

Conclusions

The three Biblical Hebrew syntactic databases present unparalleled opportunities for the retrieval of syntactic data from the Hebrew Bible and additional pre-modern Hebrew texts. While there are overlaps between the three in terms of the retrieval of

[15] Sources of synchronic variation that are often studied in modern corpus linguistics involving demographics of the speaker/writer, such as ethnicity, gender, geography (dialect) and age, cannot be clearly determined for the texts in these databases. However, a clustering of features within certain portions of the databases might lead to (or lend credence to) hypotheses about the demographics of the writer.

morphological and lexical data, each of them is different and each makes a unique contribution. A researcher working on Biblical Hebrew syntax must consult all three databases because they are built on different theoretical premises and provide different perspectives and often different kinds of data.

It is also important to view the syntactic analyses provided by the databases as *one* analysis of the data, which must be critically examined. One example can be seen in the complex phrase וְעֵץ הַדַּ֖עַת טֹ֥וב וָרָֽע (Genesis 2:9; see also Genesis 2:17 and Exodus 29:40). Each of the databases identifies הַדַּעַת as the construct noun form (even though it has the definite article) and the two adjectives that follow as the absolute forms (even though they are indefinite) within a construct phrase. Another analysis is implied in the description of many Biblical Hebrew grammars (e.g., GKC 1910, §115d; Joüon and Muraoka 2009, §124d, §124j; Brockelmann 2004, §99b; Waltke & O’Connor 1990, §10.2.2) and the lexicon entry of Köhler and Baumgartner (2001: s.v. דַּעַת)—the noun with the article is in the absolute state and not in the construct state. The items after the determined noun are indefinite accusatives.[16]

Each of the articles by the creators of the databases indicate areas for future improvement that they envision for their respective databases. All these aspirations are very welcome, especially in the expansion of the corpus to additional pre-modern Hebrew texts. We suggest in addition two other avenues of expansion. First, attention could be given to the various reading traditions of Hebrew (see recently Garr and Fassberg 2016). Second, the question of the interface of syntax with other modules of grammar such as prosody is an area that is beginning to receive attention (see Naudé and Miller-Naudé 2017; Pitcher 2017). Such an approach is part of a larger move toward complexity thinking (see Marais 2014), which has replaced reductionistic modernist and post-modernist viewpoints. The observation of Wido Van Peursen (2017, 392–394) that we are now in a phase where there is a re-unification of traditional hermeneutical and exegetical analysis of texts with the electronic discoveries of patterns in texts is part of this general trend.

Finally, we congratulate the creators of the databases for the very significant work that they have accomplished thus far and wish them every success as they continue to enhance and expand their databases in the future.

References

Adolphs, S. 2006. Introducing Electronic Text Analysis. A Practical Guide for Language and Literary Studies. London and New York: Routledge. https://doi.org/10.4324/9780203087701.

[16] We examine these constructions in a forthcoming article.

Andersen, F.I. 1970. The Hebrew Verbless Clause in the Pentateuch. JBL Monograph Series 14. Nashville: Abingdon.

Andersen, F.I. 1974. The Sentence in Biblical Hebrew. Janua Linguarum; Series Practica 231. The Hague: Mouton. https://doi.org/10.1515/9783111356808.

Andersen, F.I. and Forbes, A.D. 2012. Biblical Hebrew Grammar Visualized. Winona Lake, IN: Eisenbrauns.

Biber, D. 1990. "Methodological Issues regarding Corpus-based Analyses of Linguistic Variation." Literary and Linguistic Computing 5: 257–269. https://doi.org/10.1093/llc/5.4.257.

Biber, C., Conrad, S. and Reppen, R. 1998. Corpus Linguistics: Investigating Language Structure and Use. Cambridge Approaches to Linguistics. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511804489.

Botha, R.P. 1981. The Conduct of Linguistic Inquiry. The Hague: Mouton. https://doi.org/10.1515/9783110822946.

Bowker, L. and Pearson, J. 2002. Working with Specialized Language. A Practical Guide to using Corpora. London and New York: Routledge. https://doi.org/10.4324/9780203469255.

Brockelmann, C. 2004. Hebräische Syntax. 2nd edition. Neukirchen-Vluyn: Neukirchener Verlag.

Forbes, A.D. 2014. "A Tale of Two Sitters and a Crazy Blue Jay." In Reflections on Lexicography: Explorations in Ancient Syriac, Hebrew and Greek Sources, edited by R.A. Taylor and C.E. Morrison, 211–232. Perspectives on Linguistics and Ancient Languages 4. Piscataway, NJ: Gorgias.

Friedman, T.L. 2005. The World is Flat. The Globalized World in the Twenty-First Century. London: Penguin books.

Friedman, T.L. and Mandelbaum, M. 2011. That used to be Us. New York: Farrar, Straus and Giroux.

Garr, W.R. and Fassberg, S.E. 2016. A Handbook of Biblical Hebrew. 2 vols. Winona Lake: Eisenbrauns.

Gazdar, G. 1988. "Applicability of Indexed Grammars to Natural Languages." In Natural Language Parsing and Linguistic Theories, edited by U. Reyle and C. Rohrer. Studies in Linguistics and Philosophy, 35. Dordrecht: Springer. doi:10.1007/978-94-009-1337-0_3. https://doi.org/10.1007/978-94-009-1337-0_3.

Gazdar, G., Klein, E.H., Pullum, G.K. and Sag, I.A. 1985. Generalized Phrase Structure Grammar. Oxford: Blackwell, and Cambridge, MA: Harvard University Press.

Gesenius, W., Kautzsch, E. and Cowley, A.E. 1910. Gesenius' Hebrew Grammar. 2nd English edn. Oxford: Clarendon. [GKC]

Holmstedt, R.D. and Cook, J.A. 2018. "The Accordance Hebrew Syntactic Database Project." Journal for Semitics 27 (1): this volume.

Hughes, J.J. 1988. Bits, Bytes and Biblical Studies. Grand Rapids, MI: Academic Books/Zondervan Publishing House.

Joüon, P. and Muraoka, T. 2009. A Grammar of Biblical Hebrew. 2nd reprint of 2nd edition, with corrections. Rome: Gregorian Biblical Press.

Köhler, L. and Baumgartner, W., eds. 2001. The Hebrew and Aramaic Lexicon of the Old Testament. Leiden: Brill.

Lombard, H.C. and Naudé, J.A. 2009. "The Localisation of the EtsaTrans Translation Programme for the University of the Free State Library and Information Services." Mousaion 27: 51–74.

Marais, K. 2014. Translation Theory and Development Studies: A Complexity Theory Approach. New York and London: Routledge. https://doi.org/10.4324/9780203768280.

Marais, K. and Naudé, J.A. 2007. "Collocations in Popular Religious Literature as an instance of Language for Special Purposes: An Analysis in Corpus-based Translation Studies." Southern African Linguistics and Applied Language Studies 25, (2): 153–167. https://doi.org/10.2989/16073610709486454.

McEnery, T. and Wilson, A. 2001. Corpus Linguistics. An Introduction. Second edition. Edinburgh: Edinburgh University Press.

McEnery, T., Xiao, R. and Tono, Y. 2006. Corpus-based Language Studies. London and New York: Routledge.

Meyer, C.F. 2002. English Corpus Linguistics. An Introduction. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511606311.

Miller, C.L. 2004. "Methodological Issues in reconstructing a Language System from Epigraphic Fragments." In The Future of Biblical Archaeology: Reassessing Methods and Assumptions, edited by J.K. Hoffmeier and A. Millard, 281–305. Grand Rapids: Eerdmans.

Miller-Naudé, C.L. 2012. "Diachrony in Biblical Hebrew: Linguistic Perspectives on Change and Variation." In Diachrony in Biblical Hebrew, edited by C.L. Miller-Naudé and Z. Zevit, 3–15. Winona Lake, IN: Eisenbrauns.

Miller-Naudé, C.L. and Naudé, J.A. 2017. "A Re-examination of Grammatical Categorization in Biblical Hebrew." In From Ancient Manuscripts to Modern Dictionaries: Select Studies in Aramaic, Hebrew, and Greek, edited by T. Li and K. Dyer, 331–376. Perspectives on Linguistics and Ancient Languages 9. Piscataway, NJ: Gorgias Press.

Naudé, J.A. 2004. "Representation of Poetry in the Afrikaans Bible Translations. A Corpus-based Translation Analysis." Language Matters 35: 233–254. https://doi.org/10.1080/10228190408566214.

Naudé, J.A. 2008. "The Role of Pseudo-translations in Early Afrikaans Travel Writing. A Corpus-based Translation Analysis." Southern African Linguistics and Applied Language Studies 26 (1): 97–106. https://doi.org/10.2989/SALALS.2008.26.1.8.423.

Naudé, J.A. and Miller-Naudé, C.L. 2017. "At the Interface of Syntax and Prosody: Differentiating Left Dislocated and Tripartite Verbless Clauses in Biblical Hebrew." Stellenbosch Papers in Linguistics 48: 223–238.

Oakes, M. 1998. Statistics for Corpus Linguistics. Edinburgh: Edinburgh University Press.

Olohan, M. 2004. Introducing Corpora in Translation Studies. London and New York. https://doi.org/10.4324/9780203640005.

Pike, K.L. 1967. Language in Relation to a Unified Theory of the Structure of Human Behavior. The Hague: Mouton. https://doi.org/10.1515/9783111657158.

Pike, K.L. and Pike, E.G. 1977. Grammatical Analysis. Dallas: SIL International Publications in Linguistics.

Pitcher, S.L. 2017. "Towards the Development of an Information-based Prosodic Model for the Masoretic Cantillation Accents of Tiberian Hebrew." MA dissertation. University of the Free State.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. 1985. A Comprehensive Grammar of the English Language. London: Longman.

Schneider, W. 2007. Grammatik des biblischen Hebräisch: Ein Lehrbuch. 3rd edition. Munich: Claudius.

Sinclair, J. 1991. Corpus Collocation Concordance. Oxford: Oxford University Press.

Sinclair, J. 2003. Reading Concordances. London: Pearson Education.

Sinclair, J. 2004. Trust the Text: Language, Corpus and Discourse. London: Routledge

Snyman, C., Ehlers, L. and Naudé, J.A. 2007. "Development of the EtsaTrans Translation System Prototype and its integration into the Parnassus Meeting Administration System." Southern African Linguistics and Applied Language Studies 25 (2): 225–238. https://doi.org/10.2989/16073610709486458.

Talstra, E. 1978. "Text Grammar and Hebrew Bible I: Elements of a Theory." Bibliotheca Orientalis 35: 169–174.

Talstra, E. 1992. "Text Grammar and Biblical Hebrew: The Viewpoint of Wolfgang Schneider." Journal of Translation and Textlinguistics 5: 269–297.

Teubert, W. and Čermáková, A. 2007. Corpus Linguistics: A Short Introduction. London: Continuum.

Van Peursen, W. 2017. "New Directions in the Computational Analysis of Biblical Poetry." In Congress Volume 2016 Stellenbosch, edited by L.C. Jonker, G.R. Kotzé and C.M. Maier, 378–394. Leiden: Brill.

Waltke, B.K. and O'Connor, M. 1990. Introduction to Biblical Hebrew Syntax. Winona Lake, IN: Eisenbrauns.