Clause Structure Variation in Biblical Hebrew

Clause Structure Variation in Biblical Hebrew

A Quantitative Approach

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor of Philosophy aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. V. Subramaniam,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit Religie en Theologie
op dinsdag 1 december 2020 om 13.45 uur
in de online bijeenkomst van de universiteit,
De Boelelaan 1105

door

Martijn Naaijer

geboren te Vlissingen

Promotor:
Prof.dr. W.T. van Peursen

Copromotoren:
Dr. J.W. Dyk
Dr. D.R. Roorda

Contents

Acknowledgements

Het SynVar project was een prachtig avontuur, waarin we enorm veel nieuwe dingen gedaan hebben. Dat kan natuurlijk alleen maar bij het ETCBC.

Wido, mijn promotor, dank voor alle mogelijkheden die je me hebt gegeven in de afgelopen jaren. Eén van de mooiste hiervan is de ontwikkeling van het vak Digital Humanities and Biblical Studies, waarmee de databank van het ETCBC een onderdeel is geworden van het reguliere onderwijs. Ik denk met plezier terug aan de discussies laat op de middag op het ETCBC, waarbij wilde nieuwe ideeën met een drankje besproken werden.

Janet, mijn copromotor, jouw taalkundige expertise was van belang voor dit hele project. Dank voor het delen van je kennis op dit gebied.

Dirk, mijn andere copromotor, je had een bijzondere rol gedurende het hele project. Met het Shebanq-project kwam het principe van open data naar het ETCBC, wat een grote stap betekende voor het ETCBC en eigenlijk ook voor het hele onderzoeksveld. Zonder LAF-Fabric en Text-Fabric zou mijn proefschrift er totaal anders uit hebben gezien. Dank voor je hulp bij het gebruiken van deze tools. Het was in het begin even wennen met MQL queries in LAF, maar sindsdien zijn we een eind gekomen. Ik denk met plezier terug aan de reizen die we gemaakt gemaakt hebben. Weliswaar is de VS niet je favoriete land, maar ze hebben je tot mijn genoegen toch verbaasd met hun IPA's.

Constantijn, jouw kennis van de databank is natuurlijk onontbeerlijk voor elke onderzoeker van het ETCBC. Hartelijk dank voor je geduldige uitleg, niet alleen tijdens de "inleiding" van het SynVar project, maar ook in de daaropvolgende jaren.

Marianne en Dirk, de andere onderzoekers op het SynVar project, dank voor de inspirerende bijeenkomsten die we tijdens het SynVar project hebben gehad, waarin ieder zijn licht liet schijnen op actuele discussies in de wereld van de taalvariatie.

Robert, waar zullen we eens beginnen? Sinds onze eerste ontmoetingen in 2010 in Nijmegen hebben we veel gediscussieerd, en is je betrokken en kritische houding een inspiratie geweest voor verder kijken dan de gevestigde meningen. Dank voor de het lezen en becommentariëren van mijn teksten, en voor vriendschap en samenwerking in de afgelopen jaren.

Jarod, thanks for guiding me through Portland, our collaboration in the CACCHT project, and, of course, for your help in the very last phase of this whole process. Cheers!

Harald, dank voor de mooie bijeenkomsten in Amsterdam en Tübingen. Ik heb er niet alleen veel geleerd over statistiek en taal, maar het was ook altijd buitengewoon gezellig.

Cody, jouw komst naar Amsterdam was belangrijk voor de waardering van Text-Fabric in en buiten Amsterdam. Echter, belangrijker vind ik je vriendschap en onze discussies over onderzoek, onderzoekers en Amerikaanse politiek. It was YUGE!

Dan hebben we ook nog het gezelschap dat opereert onder de naam "Roze Theekransje". Ik ken jullie sinds mijn eerste strapatsen in Wageningen, en voor de meesten van jullie is de PhD iets uit het stenen tijdperk. Maar goed, ook mijn project is nu echt af, ik kan alvast verklappen dat het minder met planten en rare beestjes te maken heeft dan jullie projecten, tijd om daar eens op te proosten.

Het leven gaat natuurlijk niet alleen over het Bijbels Hebreeuws. Er is ook nog schaken, en als er één verbond van matadoren is dat dat standvastig weet te combineren met allerlei andere belangwekkende activiteiten als drinken en oeverloos wauwelen, dan is het wel het onvolprezen tweede team van Wageningen. Jongens, dank voor de plezierige afleiding in de afgelopen jaren, en ik kijk nu alweer uit naar de volgende match in het Bilderdijkpark.

Er is ook nog andere sport: fietsen. Dit is bijna net zo leuk als schaken, zeker als het met een (oud) schaker gebeurt. Danny, dank voor de steun en mooie tochten door inmiddels een stuk of *6* landen.

Vorig jaar heb ik een groot deel van de weekenden doorgebracht in het kantoor van Bloxs om dit werk af te maken. Collega's, die ik tijdens de crisis helaas weinig zie, hartelijk dank voor alle steun!

Christiaan, onze gedeelde interesse in discussies over diversiteit en andere onderwerpen die niets met het Hebreeuws te maken hebben zijn natuurlijk de basis voor een geweldig onderzoeksteam. Dank dat je paranimf wilt zijn. Ik kijk al uit naar jouw promotie.

Natuurlijk dank ik pater Antoine voor de gesprekken over klassieke literatuur, geloof en leven. Deze blijven een continue bron van inspiratie bij de studie.

Paul, dank dat je paranimf wilt zijn en ook voor alle schitterende avonturen door de jaren heen, waarvan de laatste in mei helaas niet doorging, maar dat gaan we binnenkort uiteraard inhalen. Ik denk ook aan Brenda, die er graag bij was geweest.

Uiteraard dank ik mijn familie, Guurt, Johan, Daniëlle, Lisette, Ursula, Erica en jullie aanhang en kinderen en katten, honden en koeien en mijn schoonfamilie, Rinus en Méry, voor jullie steun, interesse en liefde in de afgelopen jaren.

Lieve Monica, Jona en Amber, het klinkt als een cliché, maar het afmaken van een project als dit is soms best lastig, vooral voor direct betrokkenen. Dank voor alle steun en geduld in de afgelopen jaren.

Abbreviations

Abbreviations related to the Eep Talstra Centre for Bible and Computer (ETCBC) database of the Hebrew Bible

AdjP	Adjective Phrase
AdvP	Adverb Phrase
CP	Conjunction Phrase
DPrP	Demonstrative Pronoun Phrase
EPPr	Enclitic personal pronoun
InjP	Interjection Phrase
InrP	Interrogative Phrase
IPrP	Interrogative Pronoun Phrase
NegP	Negative Phrase
NP	Noun Phrase
PP	Prepositional Phrase
PPrP	Personal Pronoun Phrase
PrNP	Proper Noun Phrase
VP	Verb Phrase
carc	Clause atom relation code

Abbreviations related to the Hebrew language and the Hebrew Bible

MT	Masoretic Text
BH	Biblical Hebrew
EBH	Early Biblical Hebrew
LBH	Late Biblical Hebrew
TBH	Transitional Biblical Hebrew
QH	Qumran Hebrew
RH	Rabbinic Hebrew

Abbreviations of scholarly works

GKC Kautzsch, E., ed. and rev. *Gesenius' Hebrew Grammar*. Translated and revised by A.E. Cowley. 2nd edition, Oxford: Clarendon, 1910.

HJ Holmstedt, R., and Jones, A.R., "The Pronoun in Tripartite Verbless Clauses in Biblical Hebrew: Resumption for Left-dislocation or Pronominal Copula?", Journal of Semitic Studies LIX/1, 2014, 53–89.

JM Joüon, P. *A Grammar of Biblical Hebrew*. Translated and revised by T. Muraoka. Subsidia biblica 27. 2nd edition, Rome: Pontifical Biblical Institute, 2006.

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

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

Abbreviations of names of biblical books

Gen	Genesis
Exod	Exodus
Lev	Leviticus
Num	Numbers
Deut	Deuteronomy
Josh	Joshua
Judg	Judges
1–2 Sam	1–2 Samuel
1–2 Kgs	1–2 Kings
Isa	Isaiah
Jer	Jeremiah
Ezek	Ezekiel
Hos	Hosea
Joel	Joel
Amos	Amos
Obad	Obadiah
Jonah	Jonah
Mic	Micah
Nah	Nahum

Hab	Habakkuk
Zeph	Zephaniah
Hag	Haggai
Zech	Zechariah
Mal	Malachi
Ps	Psalms
Prov	Proverbs
Job	Job
Song	Song of Songs
Ruth	Ruth
Lam	Lamentations
Qoh	Qoheleth
Esth	Esther
Dan	Daniel
Ezra	Ezra
Neh	Nehemiah
1–2 Chr	1–2 Chronicles

Abbreviations related to statistics and machine learning

CNN	Convolutional Neural Network
GAMM	Generalized Additive Mixed Model
GAM	Generalized Additive Model
GLM	Generalized Linear Model
HMM	Hidden Markov Model
LSTM	Long Short-Term Memory
NN	Neural Network
RF	Random Forest
RNN	Recurrent Neural Network
XGBoost	Extreme Gradient Boosting

Other abbreviations

NLP	Natural Language Processing
TAM	Tense, Aspect and Mood

CHAPTER 1

Introduction

The fields of Data Science and Natural Language Processing (NLP) have made substantial advances in the last decade. Tech companies like Google provide us with tools such as translation machines and chatbots, and new developments follow each other rapidly. The techniques underlying these developments are generally based on machine learning, which presently is the most successful branch of Artificial Intelligence. With machine learning, it is possible to let computers learn from data without explicit instructions. Its success has been stimulated mainly by three factors. The development of better algorithms and better hardware are two of them, but the most important factor is the availability of large datasets on the Internet. Without data, there is no machine learning.

The relevance of these developments for Biblical Studies should not be underestimated. Various datasets containing annotated ancient texts and text corpora have been published online, and their number increases rapidly. These electronic corpora, of which many are freely available, make it possible to do large-scale analysis, in which many linguistic and other features are combined. Also, it is possible to analyze data from different languages and text corpora together using new methods and techniques. This is a huge advance over a situation in which the data are locked up in printed editions and commercial software packages without free access and possibilities to process and export the data in any desired way.

In this research, a number of new techniques are applied to the problem of linguistic variation in Biblical Hebrew (BH). Overall, BH is a relatively homogeneous language. Most of the lexicon, morphology, and syntax is identical throughout the books of the Hebrew Bible, but some linguistic variation can be observed: there is variation between genres, the Hebrew of poetic texts differs from that of prose, but also within one genre; the language of Genesis differs from that of Ezra; and even within separate books like Genesis itself there is linguistic variation.

Just like in other languages, this variation may be explained by diachronic development. Since the emergence of critical scholarship on the Hebrew Bible, most scholars have argued that much of the linguistic variation in BH is due to change over time. These scholars distinguish Early Biblical Hebrew (EBH) from Late Biblical Hebrew (LBH). EBH can be found in the Pentateuch and Former Prophets, and LBH can be found in the so-called core late books of Esther, Daniel, Ezra, Nehemiah, and

Chronicles, and maybe some other books like Qoheleth. Generally, scholars relate the Babylonian exile to the transformation from EBH to LBH, so EBH is considered to date from the pre-exilic period, and LBH from the post-exilic period.

This distinction between EBH and LBH led Avi Hurvitz to the development of a method to date biblical texts of unknown date, such as Jonah and Ruth. This method of linguistic dating has been applied widely in the field of Biblical Studies.

Next to diachronic variation, scholars have detected variation between dialects, social layers, and individual styles in the Hebrew Bible. However, finding out which of these factors play a role in which texts is easier said than done, because of various complicating factors, one being that most books have been edited, and since we do not possess the original manuscripts, it is often difficult to reconstruct the various editorial layers. The long history of transmission of the books of the Hebrew Bible may have caused both differentiation and harmonization of its language.

Besides editing and transmission, the nature of BH itself has its own peculiarities. Not only is its vocabulary relatively restricted, but due to the nature of its texts, it may be a literary language. Last but not least, the corpus of BH is limited: the Hebrew Bible contains fewer than 430,000 words. These problems, addressed by various scholars in the late 20th and early 21st centuries, led to fierce debates in the field of Biblical Studies concerning the history of BH and the possibility of dating biblical texts on the basis of their language.

The discussion as presented briefly above has led to a kind of stalemate position. The gap between those who say it is possible to date texts linguistically and those who say it is not has resulted in a situation in which scholars seem to dig in their heels.

This research is embedded in the project "Does Syntactic Variation reflect Language Change? Tracing Syntactic Diversity in Biblical Hebrew Texts".[1] With this project we try to overcome the present stalemate position. We concur with those who say that syntax may give a better impression of the history of BH than vocabulary (Polzin 1976; Joosten 2005; Rezetko 2003: 249): being more deeply rooted in the language system and being less consciously used (Henry 2004), syntax is less apt to be manipulated.

---

1 Funded by the Netherlands Organization for Scientific Research (NWO). The applicants are Dr. J.W. Dyk and Prof. Dr. W.Th. van Peursen. Next to this research on clause structure, there are subprojects on phrase structure (Marianne Kaajan) and text hierarchy (Dirk Bakker) and a synopsis of the three sub-projects (Janet Dyk and Wido van Peursen).

The main research question of this project is:

> How and to what extent does the structure of clauses in BH vary and what is the linguistic, literary, and historical background of this variation?

The subtitle of the thesis is "a quantitative approach". In studies on linguistic variation in BH, quantification of results always plays a role. For instance, a certain word or construction occurs predominantly in LBH, and a semantic alternative construction occurs predominantly in EBH. In this situation it is important to know the distribution of both the EBH and the LBH alternative throughout the Hebrew Bible, and the proportion of these alternatives in various parts of it. In traditional studies on linguistic variation in BH, quantification is dealt with only to a limited extent. The purpose of this thesis is to use methods with which a more complete quantitative impression of linguistic variation is given. Also, there is an important role for visualizing data in this thesis, because visualizations clarify more than tables of numbers. The goal of this approach is to make research on linguistic variation in BH more data-driven, and less dependent on models based on the intuition of scholars.

With the focus on quantitative methods there is less attention for conventional philology. That does not mean that traditional linguistics is not important, but the study of the history of BH does not have a strong quantitative foundation, and new technical developments justify attempting to see what they can bring to the study of ancient languages and BH in particular.

As stated above, this research is part of the Syntactic Variation project of the Eep Talstra Centre for Bible and Computer (ETCBC), in which syntactic variation in BH is studied on three different levels. Marianne Kaajan describes phrase structure variation,[2] my research addresses clause structures, and Dirk Bakker investigates larger text units.[3]

The structure of the thesis is as follows:

Chapter 2 is a general review of the literature on linguistic variation in BH. In each of the subsequent chapters, there is a review of the topic discussed in it.

Chapter 3 contains a description of the scientific framework and the data, more specifically of the ETCBC database from which the data in this research were extracted.

---

2 Kaajan, M.E., Syntactic Variation in Non-Verbal Phrase Structure in Biblical Hebrew (in preparation).

3 Bakker, D., Syntactic Variation in Classical Hebrew: A Text-Grammatical Approach (in preparation).

Because a clause has as core either an expression of "being", without or with the verb "to be", or a verb with its satellites, these two types of clause structure are dealt with in chapters 4 and 5.

Chapter 4 contains an analysis of various expressions for "to be" in BH. After the review in section 1, section 2 describes the variation between clauses with a subject and predicate complement with and without היה, using a Generalized Additive Mixed Model (GAMM). In section 3 the use of the particle יש is analyzed by contrasting clauses with and without יש. In the analysis, linguistic variation is studied using Random Forest and Extreme Gradient Boosting. Section 4 also uses Random Forest and Extreme Gradient Boosting to study the tripartite verbless clause.

As an example of the other type of clause structure, chapter 5 deals with verbal valency, focusing on complex verbal valency patterns. The case study is נתן and שים with double object constructions.

In chapter 6, clause structure variation is approached from a different perspective. Instead of focusing on a particular syntactic feature, all clauses from a relevant corpus are studied and a model is trained on these clauses in order to predict their language class. In this analysis, a Long Short-Term Memory (LSTM) network is used.

The approach in all the chapters is that first a dataset is generated with Text-Fabric,[4] a Python API of the Hebrew Bible (Roorda, Kalkman, Naaijer, and Van Cranenburgh 2014; Roorda 2018), then the data are explored and existing hypotheses about the structure of linguistic variation are investigated and tested using a number of statistical and machine learning techniques.

For those who are interested in more detail on the techniques used in the thesis, I have included appendices with technical details. The quantitative work is completely reproducible, the data can be downloaded from GitHub (see https://github.com/Dans-labs/text-fabric for instructions), as can all the Python and R scripts for preprocessing and analyzing the data (https://github.com/martijnnaaijer/phdthesis).

4 Except chapter 5 on verbal valency, in which most of the data was prepared manually.

CHAPTER 2

Review of history of scholarship

2.1. Introduction

Most of the important modern standard reference works on the history of BH accept the division of the history of BH into three periods:[1] Archaic Biblical Hebrew (ABH), Early Biblical Hebrew[2] (EBH), and Late Biblical Hebrew (LBH).

According to these works, ABH can be found mainly in a number of poetic texts in the Pentateuch and Former Prophets: Gen 49 (the blessings of Jacob), Exod 15 (the Song at the Sea), Num 23–24 (the oracles of Balaam), Deut 32–33 (the prayer and blessing of Moses), Judg 5 (the song of Deborah), 1 Sam 2:1–10 (the prayer of Hannah), Hab 3, and various Psalms, including Ps 18 // 2 Sam 22 and Ps 68.[3] These poems show a concentration of linguistic features found in lower concentrations elsewhere in the Hebrew Bible.[4] The linguistic similarities between the ABH texts, the Ugaritic corpus, and the Amarna letters led scholars like Cross and Freedman to state that these texts contain the earliest phase of BH, which can be dated to the pre-monarchic and early monarchic period.[5] Extra-linguistic arguments play a role as well in dating separate texts: Exod 15 could have been influenced by Ugaritic mythology, as suggested by Cross (1994: 113–144). Because of the limited size of the ABH corpus, it will play only a minor role in this study.

---

1 This divison is accepted by Kutscher (1982), Sáenz-Badillos (1993), and the *Encyclopedia of Hebrew Language and Linguistics* (Khan, ed. 2013, in the lemma: "Biblical Hebrew: Periodization" by Hornkohl).

2 Also called Standard Biblical Hebrew (SBH) or Classical Biblical Hebrew (CBH).

3 This list is from Mandell (2013). There are varying opinions on which texts are characteristic of the ABH corpus and which texts show ABH in its "purest" form; for this issue see Robertson (1972).

4 For a description of various features, see for instance Cross and Freedman (1975); Sáenz-Badillos (1993: 56–62); Vern (2011). A description of the verb system of ABH can be found most comprehensively in Notarius (2013).

5 Cross and Freedman (1994: ch 1). Sáenz-Badillos (1993: 56) does not date the ABH texts explicitly to the pre-monarchic period, he simply classifies ABH under Pre-Exilic Hebrew. For a critique of assigning an early date to the ABH texts, see Vern (2011). For further recent discussions, see Pat-El (2014) and Notarius (2015). For a critique of the methodology of Cross and Freedman (1997, reprint), see Goodwin (1969). Recently the issue of ABH was discussed extensively in the introduction and three articles in Barmash, ed. (2017: 47–118).

Much more material is available for the study of the later phases of BH. Various 19th century scholars observed the differences between the Hebrew of the Pentateuch and Former Prophets on the one hand, and of the undisputed late books of Esther, Daniel, Ezra, Nehemiah, and Chronicles on the other hand.[6] According to Gesenius, most of the Pentateuch and Former Prophets were written in the pre-exilic period, but later redaction must be supposed (Gesenius 1815: 23).[7] He also stated that other early texts include Psalms, Proverbs, Job, and the prophetic books of Amos, Hosea, Micah, and Isaiah (Gesenius 1815: 24). Gesenius gives several linguistic characteristics of early prose and early poetry, but he generally judges prophetic literature on the basis of style. On Ezekiel he says: "*Zwar der originellste der Dichter, dessen üppige Phantasie in neuen gigantisch-grotesken Bildern schwelgt, besitzt er doch zu wenig Geschmack und Concinnität, um den Namen eines classischen Schriftstellers zu verdienen*" (Gesenius 1815: 25). Gesenius postulated that the linguistic differences can be explained by assuming that during the exile, when there was increased contact between Hebrew and Aramaic speakers, the Hebrew language was influenced by Aramaic. Throughout the 19th and 20th century the basic framework of Gesenius' ideas, consisting of an early and a late phase of BH separated by the exile, would remain influential.

After a period of consensus[8] on the diachrony of BH, a number of works were published in the 1990s in which the foundations of the chronological model were questioned. Cryer (1994) asks whether any substantial variation at all can be traced in BH, thereby questioning the distinction between EBH and LBH, Davies (1995) argues that it is possible that books written in EBH could originate in the post-exilic period, and, according to Knauf (1990), BH is a rather artificial language.

Although the idea that BH, just like every other natural language, has developed over time is reasonable, there are various factors that make it difficult to undertake a description of this development. Most investigations of BH are based on the language of the text in the Codex Leningradensis (L), which was produced in 1008/1009 CE. This manuscript is the oldest complete manuscript of the Hebrew Bible. Slightly older than Leningradensis is Codex Aleppo (late 10th century CE), which was partly destroyed in a fire in 1948. These two codices are representatives of the so-called Masoretic Text or MT. Although the term Masoretic Text suggests that it is a specific text, according

6 Although the variation in BH and the influence of Aramaic on BH in later compositions was observed earlier by Hugo de Groot (Hornkohl 2014: 2–3).

7 In n. 24 Gesenius indicates that Deut 33:7 must have been written during the exile.

8 A good overview of this consensus can be found in Sáenz-Badillos (1993).

to Tov it would be better to say Masoretic Texts or M-group of texts, because there are many differences among the Masoretic manuscripts (Tov 2012: 25). The Masoretic Text tradition goes back to older traditions, which may go back to the Second Temple period (Tov 2012: 24). This so-called Proto-Masoretic Text can be found in various Dead Sea Scrolls (DSS),[9] and is probably the Vorlage of several translations, like the Peshitta, the Targum and the Vulgate. This does not necessarily mean that the MT is the oldest text of the Hebrew Bible. Some books in the Septuagint differ strongly from the MT and are apparently based on a different Hebrew text, which may be older than the MT.

The second most important group of biblical manuscripts is the biblical DSS. Among the DSS, the oldest biblical manuscripts of substantial length can be found. These scrolls were produced between the 3rd century BCE and the 1st century CE. Some manuscripts have a text that is similar to the MT, while other manuscripts have texts that deviate substantially from the MT.[10]

2.2. Dating biblical texts: Avi Hurvitz

A major advance in the research on variation in BH was achieved by Avi Hurvitz. In some influential works[11] he argued that by using a straightforward method it is possible to distinguish between EBH and LBH,[12] and that based on this difference, it is possible to linguistically date biblical texts of unknown date. These texts used to be dated by theological, historical, and literary criteria, but, according to Hurvitz, using language as the dating criterion is more objective than the criteria of so-called Higher Criticism (Hurvitz 1973: 74). By counting the number of late linguistic features of a specific biblical text, Hurvitz was able to date several texts as late, including the prose-tale of Job and several Psalms (Hurvitz 1973 and 1974).

9 An example is MasLev^{b}, Tov (2012: 29 n. 8).

10 Strong deviations from L can be found in for instance 1QIsa^{a}.

11 Hurvitz (1974) on the book of Job, and Hurvitz (1972) on the controversial issues of the language and date of P.

12 Hurvitz and many others think that EBH can be found mainly in the Pentateuch and Former Prophets and that LBH can be found mainly in Qoheleth, Esther, Daniel, Ezra-Nehemiah and Chronicles.

2.3. Criteria of linguistic dating

According to Hurvitz, dating texts is not without complications. A post-exilic author can write in an archaizing style, so it may be difficult to distinguish between real EBH and late language that is archaized, but an early author could not write in a post-exilic style. Therefore, if late linguistic features are found in a text, it can be dated to the post-exilic period, but the absence of late language cannot automatically lead to the conclusion that a certain text is pre-exilic (Hurvitz 1973: 75).[13] Due to the relative uniformity of BH and an archaizing style of several texts, it is not possible to be more specific about the date of a text other than to say that it is post-exilic.[14] Hurvitz makes use of the MT without textual emendations (Hurvitz 1973: 74).

To be able to date a text as late, Hurvitz uses four criteria. With the first three criteria, one can isolate separate late linguistic features, the fourth criterion states that a text is late if there is an accumulation of these late features in the text. The first criterion for isolating a late feature is that of distribution: a feature must occur predominantly or exclusively in the core LBH books.[15] The second criterion is that of opposition or contrast. For every late feature, there must be an early equivalent, which occurs predominantly or exclusively in the EBH books. With this criterion one can guarantee that it is not a coincidence that a late feature occurs in some books and not in others, since an early feature has been replaced by a late one.

According to the third criterion the late feature should occur in late extra-biblical Hebrew or in late texts in cognate languages. The main sources are the DSS and Rabbinic texts (mainly the Mishnah), but some authors also search for features in the Targumim. With this third criterion it can be guaranteed that a late feature is not just typical of the personal style of a post-exilic author, but that it had a wider use. With these criteria, late features can be extracted from biblical texts.

One or two late features can occur in a text by coincidence, according to Hurvitz, so to be able to assign a late date to a text the fourth criterion states that there must

13 This view seems to be corrected by Hendel and Joosten (2018: 65–70), who point at positive traits of EBH and linguistic relationships between epigraphic Hebrew and EBH.

14 Concerning the prose tale of Job, Hurvitz (1974: 31) says: "That is to say, we may actually be dealing not with an archaic but rather with an archaizing language. Another possible explanation of the existence of old linguistic elements could be, of course, that (some of) the material in the Prose Tale is indeed old, its final form being shaped, however, in a late period".

15 Hurvitz (1973: 75) is not precise in specifying the core late corpus, but he does include Qoheleth. Some others, for instance Bergey (1983), do not include this book, although he does not explain why he does not include it.

be an accumulation of late features in a text.[16] Every late linguistic feature gets a weight of one, no matter how often it occurs in a late text. Hurvitz is not precise about what number of features can be called an accumulation, except that it must be more than the already mentioned one or two. Hurvitz (1974) describes seven late linguistic features in the prose tale of Job in a stretch of text of 1070 words,[17] so apparently this is enough to assign a text to the post-exilic period.

2.4. Transitions: from EBH to LBH

According to Gesenius, the earlier form of BH was in use before the exile and the later form started to be used during the exile, due to Aramaic influence. Gesenius calls the later form spätere chaldäisch-gefärbte Sprache, which can be found in the books of Jonah, Qoheleth, Esther, Daniel, Chronicles, and some Psalms. Gesenius detected the Aramaic influence in the vocabulary, morphology, phrases, orthography, and syntax (Gesenius 1815: 28–30). He thought that the language of other late books like Zechariah, Malachi, the Song of Songs, Ezra, and Nehemiah was a bit purer (*reiner*)[18] (Gesenius 1815: 27).

Later in the 19th century, S.R. Driver closely followed these ideas of Gesenius. In his "Introduction" (Driver 1892b) he described the time of the early biblical stories as the "golden age" of Hebrew literature and its language as the "purest form" of BH (Driver 1892b: 473). This language can be found in the Pentateuchal layers J and E, the books of Samuel and Kings and the early stories in Judges. He called the later syntax "labored", "inelegant", and "deteriorated"[19] (Driver 1892b: 473–474).

An interesting question is when the Hebrew language started to shift from EBH to LBH and in which books this transition is visible. Driver argued that the transition took place in the time of Nehemiah (5th century BCE, Driver 1892b: 148), because in the early post-exilic books, he did not find the late characteristics (Driver 1892b: 336). Among these are the books of Haggai, Zechariah, and Malachi, and the Priestly

16 A description of the criteria of dating BH texts can be found in Hurvitz (1973: 74–76), but also in many other works of Hurvitz and others, for instance, in Hurvitz (2012: 267–268); Hurvitz (2014: 9–11); Bergey (1983: 16–21), and adapted for studying dialectal variation in Rendsburg (1990: 15–17).

17 Job 1, 2 and 42:6–17, for a count of the words, see https://github.com/MartijnNaaijer/phdthesis/blob/master/Ch2_History_of_scholarship/word_count_prose_tale_job.ipynb.

18 For Gesenius, the later language was clearly in a state of decay; therefore, he calls the language of the books of Haggai, Malachi and some later Psalms "*kraftlos und wässerig*", p. 26.

19 The late language can be found mainly in Qoheleth, Esther, Daniel, Ezra, Nehemiah, and Chronicles.

layer of the Pentateuch, which was dated to the post-exilic area by Driver, on the basis of non-linguistic grounds. Several later authors agree with Driver that the late language is absent in these books,[20] although Rendsburg argues that it would be better to categorize Haggai among the core LBH books, based on the attestation of an accumulation of LBH features in this book (Rendsburg 2012b). This means that there is no disagreement on the date of these books: there is disagreement on the linguistic profile of the book of Haggai and to a lesser extent the book of Zechariah.

The language phase between EBH and LBH is often called Transitional Biblical Hebrew (TBH). It is found chronologically and typologically between EBH and LBH (Young, Rezetko, and Ehrensvärd 2008, volume 1: 51), and it has a mixture of EBH and LBH linguistic features.

Various texts and books have been described as being written in TBH: the Priestly Source, Isa 40–55, Jeremiah (Hornkohl 2014), Ezekiel, Jonah, Ruth, and Lamentations.[21] The question as to what is the best representative of TBH is dealt with in Polzin (1976), Hurvitz (1982), and Rooker (1990). Polzin (1976: 112) concludes that P is written in TBH and suggests that Ps (the *Sekundäre Priesterschrift*) is typologically later than Pg (the *Priesterliche Grundschrift*). Hurvitz (1982) criticizes Polzin and finds Ezekiel the best exemplar of TBH, as does Rooker (1990). In Hurvitz's opinion, P is written in EBH, so it must be early and the exilic book Ezekiel shows clear traces of what would develop further into LBH.[22] This means that for Hurvitz the period of transition starts in the 6th century BCE instead of in the 5th century, as Driver thought. Polzin and Hurvitz have differing opinions, not only on what is the best exemplar of TBH, but also on the presuppositions and methodology of their research. Polzin (1976: 124) studies syntactic features, which he finds a more reliable basis for dating texts than lexicon. According to Hurvitz, both grammar and lexicon can be used. Other differences are that Polzin prefers to use non-synoptic Chronicles instead of the whole book. For him, the other best examples of LBH are Ezra and the non-memoir sections of Nehemiah, whereas Hurvitz uses all the core LBH books as good examples of LBH, although often they tend to archaize. Polzin (1976: 4) sees the 6th century inscriptions as "late Hebrew sources"; for Hurvitz (1982: 162 n. 20) they are sources of EBH,[23] and for Hendel and

20 Ehrensvärd (2003: 175–187); Young, Rezetko, and Ehrensvärd (2008, volume 2: 47). Shin (2007) argues that Zechariah and Malachi contain several late lexical features.

21 See Young, Rezetko, and Ehrensvärd (2008, volume 1: 51–52), for further references.

22 Hurvitz (1982), Appendix 1 and 2.

23 Similarties and differences between the works of Hurvitz and Polzin are discussed in Young, Rezetko, and Ehrensvärd (2008, volume 1: 25–27).

Joosten (2018: 71–72) epigraphic Hebrew is close to EBH, but it also shows signs of LBH, maybe because most of the Judean inscriptions are relatively late.

2.5. Transitions: from LBH to QH and RH

Since their discovery, the DSS have shed light on many aspects of the history of BH, due to the fact that the scrolls can be dated relatively precisely (3rd century BCE–1st century CE). However, their language has led to a number of new questions. The corpus of the DSS is often treated as a unit, but there is substantial linguistic variation between the scrolls.[24]

In many scrolls, several linguistic features can be distinguished that can be related to various traditions of Hebrew, so the development of Qumran Hebrew (QH) in relation to the earlier BH and the later Rabbinic Hebrew (RH) does not seem to be a linear process (Baasten 2006: 3). Morag (1988: 149) distinguishes between General Qumran Hebrew (GCH), Copper Scroll Hebrew (CSH), and Qumran Mishnaic (QM). General Qumran Hebrew can be recognized by the presence of the long form of the personal pronouns הואה and היאה and the verbal form יקוטלהו (imperfect 3ms singular with pronominal suffix).[25] Copper Scroll Hebrew distinguishes itself by the use of the relative -ש instead of אשר and the ין plural morpheme; Qumran Mishnaic shares many features with RH, although it has its own distinctive features (Morag 1988: 149).

The question of whether QH was a spoken or a literary variety of Hebrew has been answered in various ways. Sáenz-Badillos (1993: 131) follows Goshen-Gottstein in his opinion that the members of the sect tried to remove all colloquial characteristics, trying to recreate a purely literary, biblical form of Hebrew, which reflects their religious "purity". Blau's position resembles this opinion. According to him QH is close to BH, although it has been influenced by vernaculars like Aramaic and an early form of RH, called Middle Hebrew (Blau 2000: 25)

According to Qimron and Morag, the grammar of QH shows clear traits of a spoken language. According to Morag (1988: 163), several of the features he isolated in QH have their origin in "living, spoken language". According to Qimron (2000: 244), Tiberian Hebrew has too often been equated with BH in general, and everything else as a

24 Baasten (2006: 2–8) and Sáenz-Badillos (1993: 130–147) give an overview of the various opinions on linguistic variation in the DSS.

25 Morag (1988: 151–152) discusses ten non-lexical linguistic features characteristic of GCH.

deviation from that standard. He thinks that most of the consonantal skeleton of the MT and the DSS reflects the Jerusalem dialect, and that the Tiberian vocalization reflects another dialect.

Hurvitz (2000b: 113) has a middle position in this debate. For him, QH is a mixed language:

> This means, as already indicated, that we are dealing here with a composite language, whose 'filters' secured for future generations all kinds of linguistic ingredients—biblical and extra-biblical, Jerusalemite and peripheral, written (standard) as well as spoken (non-standard).

From the perspective of sociolinguistics, Schniedewind argues that QH was an anti-language, which uses specific vocabulary for certain central activities and beliefs in the subculture (Schniedewind 2013: 177–178).

From a purely chronological perspective, QH is younger than BH. For Hurvitz, QH is a continuation of LBH, which is in line with his method of linguistic dating. Together with RH, the language of the DSS forms the most important point of reference of post-exilic Hebrew in the criterion of extra-biblical attestation. Blau agrees with this vision. For him, QH is not only a literary language, it is also the latest stage of the biblical language (Blau 2000:25). For others, QH cannot be seen as a direct continuation of LBH, because it has its own specific peculiarities that cannot have been developed from LBH (Sáenz-Badillos 1993: 132). Young also dismisses the view that all of QH is closely related to LBH, because it has its own characteristics. He argues that some QH texts are more or less indistinguishable from EBH. According to Young, the concentration of late features in Pesher Habakkuk is comparable with the concentration in EBH, and, therefore, its language is more closely related to EBH, which he sees as a conservative style of Hebrew, than to LBH (Young 2008; Hendel and Joosten 2018: 94).

Another central problem in the debate is how BH relates to RH. Concerning RH there is general consensus that it deviates strongly from BH (Young, Rezetko, and Ehrensvärd 2008, volume 1: 223; Pérez Fernandez 1999: 2). Whether this is due to the fact that it is a spoken variety of Hebrew (Pérez Fernandez 1999: 2) or is a later form of Hebrew, or both, is an open question. In the Hurvitzian tradition RH is always used as one of the benchmarks of late Hebrew, whereas others argue that some (proto-)form of RH may have co-existed with BH even in the pre-exilic period (Rendsburg 1990a; Young 1993; Young, Rezetko, and Ehrensvärd 2008, volume 1: 241–243)

2.6. Other explanations of linguistic variation

The diachronic interpretation of linguistic variation in BH clearly dominates the scholarly literature, but a range of alternative explanations have been proposed.

In the past decades, Gary Rendsburg has published extensively on regional variation in BH (for instance, Rendsburg 1990b and 2002a; Noegel and Rendsburg 2009), for which he uses Hurvitz's method with some modifications. According to Rendsburg one can distinguish between biblical texts that were written in Judean Hebrew and non-Judean or Israelian Hebrew. Approximately three quarters of the Hebrew Bible was written in Judean Hebrew (Rendsburg 2003: 9), but clear traces of other dialects can be found in books like Proverbs, Job, Qoheleth, and the stories of the northern prophets and kings in the book of Kings (Rendsburg 1990b: 8–13).

In "Diglossia in Ancient Hebrew", Rendsburg (1990a) tries to find traces of diglossia in BH. Most of the Hebrew Bible is written in a formal, literary language, but there may be traces of a colloquial variety. This disctinction between formal and colloquial language can be seen clearly in a variety of Semitic languages, and it is visible also in post-BH: QH is formal and RH is a more colloquial variety. Rendsburg studies a number of mainly morpho-syntactic features to show that diglossia exists in the Hebrew Bible, like gender neutralization (generally using masculine forms, where the standard language distinguishes between masculine and feminine forms, e.g., 2mp אתם and 3mp הם instead of 2fp אתן and 3fp הנה) and incongruence (gender and/or number do not agree, for instance, between subject and predicate within a clause).

Frank Polak has made a number of contributions from the perspective of sociolinguistics. Polak describes it as "the interaction between the way language is used in communication, the social conditions surrounding the communication process, and the speaker's attitudes to this process" (Polak 2006: 589). In a long article, Polak (2003) distinguishes different styles in BH. In the Saul-David narratives and many prophetic legends, one can find narrative texts with short, simple clauses in parataxis, which probably goes back to spoken language and oral narrative, whereas in other books, mainly Kings and post-exilic literature, one can find narrative texts with intricate sentence structure, long noun phrases, and subordinate clauses. Such a style requires more planning and likely goes back to a scribal chancery (Polak 2003: 38–39). Polak calls these two styles the rhythmic-verbal and complex nominal style. The rhythmic-verbal style is related to oral traditions: it has short sentences in which the verbs generally have no or only one argument. The complex nominal style is characterized by longer clauses, often with more than one argument per verb. Polak substantiates

his claims by a number of examples[26] and states that the rhythmic verbal style fits in a culture in which only a few people knew how to write (Polak 2003: 66). The complex-nominal style can also be observed in Aramaic legal documents from the Persian period. It is likely that the culture of the Aramaic scribal chancery influenced scribal practice in Hebrew (Polak 2006a: 46 and the following pages), so the difference in style reflects a cultural difference.

The persistent influence of Aramaic on LBH was due to the political function of Aramaic in the post-exilic period when it became a prestige language (Polak 2006b: 592). Polak distinguishes, for instance, between eastern official Aramaic syntax in infinitive clauses in embedded letters in the book of Ezra and western syntax in similar clauses in the narrator's text. This difference is explained by Polak as due to the official nature of the letters versus the more colloquial style when the author of Ezra addresses his Judean audience (Polak 2006b: 595–596). The influence of Aramaic on LBH probably came via Official Aramaic, because a substantial amount of late vocabulary is related to administration or commercial purposes (Polak 2006b: 597).

2.7. Critique of linguistic dating

The way Hurvitz's method of linguistic dating has been applied, both by himself and by others, has generated critique. According to some it is possible to date biblical texts linguistically, but in some cases, there is a better explanation of the particular linguistic features, and according to others, linguistic dating of biblical texts is not possible at all.

One of the main representatives of the first opinion is Rendsburg. In various publications (e.g., Rendsburg 2002b, 2003b, 2012a) he gives a different interpretation of the features that are considered to be characteristics of the post-exilic language by some scholars, for instance, in the case of supposed Aramaic influence on the Hebrew language. Several of the features typical of ABH are Aramaic-like features, but despite this, these texts are not dated as post-exilic by those who think that the eccentricities of the language of the archaic poems give an indication of their date.[27] Rendsburg indicates that the supposed Aramaic influence in the book of Job can be explained

26 Among which are the parallel passages of the appointment of the elders in rhythmic-verbal style in Num 11:16–17, 24–30 and in complex-nominal style in Deut 1:9–17 (Polak 2006a: 67–70).

27 Although according to some these Aramaic-like features are an indication of a late date, for instance, in the case of Exod 15; cf. Noth (1959: 98).

better by assuming style switching. The speakers in the book of Job are portrayed as foreigners by letting them speak in a foreign manner (Rendsburg 1991: 92). Other texts for which he supposes style switching are the Balaam Oracles in Num 23–24, the Dumah Oracle in Isa 21:11–12, the Massa poetry in Prov 30–31, and in the conversation of Laban with Jacob and Jacob with his wives in the book of Genesis (Rendsburg 1991: 92–95). Most scholars have interpretated the Aramaic features in the Song of Songs as a sign of its post-exilic date. According to Rendsburg, this is a misinterpretation: these Aramaisms should be seen as features of Northern or Israelian Hebrew (Noegel and Rendsburg 2009: 8).[28] Other, Mishnah-like features, which are interpreted as signs of late language, are explained by Rendsburg as colloquialisms (Rendsburg 1991: 83–87). The issue of misinterpretations of Aramaic-like features is discussed extensively by Rendsburg (2002b, 2003b).

A more principled critique of linguistic dating of biblical texts is given by Young, Rezetko, and Ehrensvärd (2008). In the first volume of this work, the authors present their critique, in the second volume they include, among other things, an alternative explanation for the linguistic variation in BH, and a list of features detected by other scholars as characteristics of LBH. Volume one starts with a description of the work of various authors who have written on linguistic dating, especially Avi Hurvitz. His method is described in detail in chapter 2 and in subsequent chapters this method is applied consistently to show the problems of linguistic dating.

Their critique of linguistic dating has several components:

- Exegetes argue that most of the books of the Hebrew Bible had a long and complex history of editing. If this is true, it is strange that this is not visible in the language of the redactional layers. The same can be said of the long pre- and post-exilic periods. Why is there no clear linguistic development visible within the literature conventionally assigned to each of these periods (Young, Rezetko, and Ehrensvärd 2008 volume 1: 57–58)?
- The Hebrew Bible has a long history of transmission. It is not clear how conservatively the text of the Hebrew Bible was transmitted. In most studies concerning

28 Other texts for which Rendsburg suspects a northern provenance are the blessings to the northern tribes in Gen 49, Deut 32, Deut 33:10, the stories of the northern judges in the book of Judges, the stories of the northern prophets and kings in the book of Kings, Isa 24–27, Hosea, Amos, Mic 6–7, selected Psalms (the largest collections are the Asaph and Korah collections), Proverbs, Qohelet, Song of Songs, and Neh 9 (Noegel and Rendsburg 2009: 4–5). For a list of Israelian Hebrew features, see Rendsburg (2003a).

linguistic dating, the MT is the single point of reference, but according to Young, Rezetko, and Ehrensvärd, there is no a priori reason why the MT should be favored over other traditions, like the Samaritan Pentateuch and the DSS. In Rezetko and Young (2014), Appendix 2, the authors collect all the variants of the MT Samuel and the DSS containing portions of Samuel and conclude that there is no specific direction of variation. In this, their results deviate from Kutscher's (1974), who argued that the language of the MT of Isaiah is older than that of 1QIsa[a], although the manuscript of 1QIsa[a] predates the Codex Leningradensis by a millennium.

– An often-recurring issue in linguistic dating is literary-linguistic circularity. Of most of the extant EBH texts we cannot be sure that they were written before the exile, with the exception of the early inscriptions. This means that for nearly all the biblical texts written in EBH only guesses can be made concerning their date of origin. In some works, scholars argue that these texts are early because there is a lack of late features. However, we know that specific features are late, because they can be contrasted with alternatives in the EBH books, leading to the problem of circularity: for determining the date of texts or books, one depends on the date of linguistic features, and for determining the date of the linguistic features one depends on the date of texts or books.

From a quantitative perspective, chapter 5 of volume one gives the most interesting argument. In this chapter on the accumulation of late features, a number of 500-word[29] samples of texts from EBH, LBH, Ben Sira, the Arad Ostraca, and the DSS are selected. In these samples, the number of late features are counted. A comparison of the resulting numbers shows that the number of LBH features is highest in samples from the core LBH books, as expected. A low accumulation of late features can be found in several of the EBH samples, but there is also a number of surprising observations (Young, Rezetko and Ehrensvärd 2008, volume 1: 132–139). In the case of the synoptic passages 1 Kgs 22:6–35//2 Chr 18:5–34, the core LBH Chronicles displays a higher accumulation of late features than EBH Kings. Among other remarkable results are the low accumulation in post-exilic Zechariah, the prose tale of Job, and a relatively high accumulation in the Arad Ostraca. This approach of evaluating accumulation of features was criticized by Hornkohl (2014: 37–41) for several reasons, one of which being the way individual features are weighted. Instead of using Hurvitz's approach of giving every feature a weight of one, he proposes to let the weight of the feature

29 More precisely, 500 graphical units, so בראשית is counted as one.

depend on the frequency in the text. This is reasonable, of course, but it also means a substantial methodological change relative to earlier research. Finding out which weight a feature should be given is a whole area of research in the field of statistics of language. A common approach is to use tf-idf.[30] Likewise, Hornkohl (2014: 39) criticizes the sample size of 500. He argues that the samples should be larger. Larger samples likely produce more reliable results than small samples, of course, but similar to the critique on the weight of individual features, this could mean a substantial change in the results of existing studies on linguistic dating. Both the critique of Young, Rezetko and Ehrensvärd on the traditional approach as well as Hornkohl's critique on Young, Rezetko and Ehrensvärd make sense. It makes one wonder why these points have not been adopted in scholarly practice yet.

In the second volume of Young, Rezetko, and Ehrensvärd (2008), the authors argue that a better explanation of the data is that there are roughly two styles in the BH narratives, coinciding with EBH and LBH. In this model EBH is a conservative kind of BH and LBH is freer: there is simply more linguistic variation within and between the LBH books than within and between the EBH texts. Throughout the two volumes the authors keep using the terms EBH and LBH, but in their "new synthesis" they use it without the chronological significance that these terms have in the diachronic model. According to this synthesis, EBH and LBH were used both before and after the exile, though in their view the evidence is stronger for post-exilic than pre-exilic LBH (e.g., possibly Qoheleth).

Young, Rezetko, and Ehrensvärd (2008) dealt with linguistic dating proper, while in Rezetko and Young (2014) the view is broader, focusing on historical linguistics in relation to BH. To avoid confusion, in this work the authors use the terms Standard Classical Hebrew (SCH) and Peripheral Classical Hebrew (PCH) instead of EBH and LBH. Still, the authors do not think that linguistic dating is a fruitful enterprise, in historical linguistics in general it is an uncommon activity. They want to integrate methods from historical linguistics and textual and literary criticism with Hebrew linguistics to lay a new foundation for Hebrew historical linguistics. This means that their approach is not anti-diachronic (see also Rezetko 2013: 68); rather, they only argue against linguistic dating as a central activity in the historical linguistics of the Hebrew language. In a long report, they develop their position further (Rezetko and Young 2019).

30 This stands for "term frequency—inverse document frequency". See for explanation and some simple experiments with R: https://cran.r-project.org/web/packages/tidytext/vignettes/tf_idf.html.

A similar critique can be found in Blum (2016). According to Blum, language is not a more decisive criterion of dating texts than other criteria (Blum 2016: 303). Similar to Young, Rezetko, and Ehrensvärd (2008), Blum argues that it is not very likely that the Pentateuch and the Former Prophets as a whole date back to the First Temple Period. He follows Wellhausen and others in their opinion that P is exilic, instead of pre-exilic as argued by Hurvitz, and that the formation of P was a long process involving multiple authors, who tried to make it like a kind of "Mosaic epigraphy" (Blum 2016: 314). As a case study, Blum discusses Gen 15 and 24. He starts by referring to a study by Rofé (1990) on Gen 24, according to whom this is a post-exilic text. He bases his opinion on various different arguments.

In the first place, he points to various linguistic links with late literature, for instance the use of אלהי השמים (Gen 24:3–7), which occurs elsewhere only in Jonah, Ezra, Nehemiah, and Chronicles (Blum 2016: 315). Other arguments for a late provenance of Gen 24 are the structure (prayer - God's answer - thanksgiving), which can be found also in other late literature (Daniel, Judith, and Tobith; Blum 2016: 316) and, the most important argument for a late date of the text is the pragmatics of the text. In Gen 24, a conflict is discussed between marrying a woman from one's own group and staying in the promised land. The conflict arises whether one should leave the land in order to find the right woman. In that case, staying in the promised land is more important. According to Rofé and Blum this issue is directed to the audience of the text and fits well in the context of Persian Judah or later (Blum 2016: 317–318). Finally, there are various intertextual relationships with other texts that are apparently presupposed by the story of Gen 24, namely, the Abraham stories, the Jacob story and the surrounding text of P (Blum 2016: 318–319).

An interesting case study showing the difficulties of linguistic dating is offered by the various opinions on the language and date of Second and Third Isaiah (chapters 40–66). Since the emergence of critical scholarship, it has been stressed that the book of Isaiah has a long and complicated history of composition and editing, and, on the basis of non-linguistic arguments, it was proposed that the book of Isaiah falls roughly into three parts, of which the second and third parts are exilic and post-exilic, respectively. According to most exegetes and linguists the language of these chapters is close to EBH (Driver 1892b: 473). However, in a short article Shalom Paul provides a substantial list of innovations from grammar and the lexicon of BH that can be found in Isa 40–66, and these innovations are relevant for the date of these chapters (Paul 2012, see also support for this position in Rendsburg 2012b: 330 n. 6). Whereas generally it is argued that the language of Isa 40–66 is early or close to early, according to Paul the language shows clear signs of later development.

These, however, are not the only possible positions. In a collection of essays of various conservative scholars, Rooker argues that the book of Isaiah is a unity on the basis of its language (Rooker 2015). Although he is aware that it will not convince many people that the whole book of Isaiah was written in the eighth century by the prophet Isaiah, he argues that the language of the Isa 40–66 is not late, but early, resembling the language of Isa 1–39. He does so by showing that the features that are shown to be characteristic of LBH often are not as typical of LBH as Paul wants his readers to believe. In some cases, other factors explain the data better. For instance, in the case of the pronominal object suffix directly attached to a verb, it is more likely that the poetic character of Isa 40–66 plays a role than that it is a sign of linguistic development (2015: 203–204). In most cases, however, the distribution of the features described by Paul do not meet the criterion of distribution, according to Rooker (2015: 221). Several of the features described by Paul do not occur in the core late books, but are hapaxes in the book of Isaiah (Rooker 2015: 221). In itself, it is true that those features occur in late literature,[31] but their absence from the core late books is a sign that these "late" features are not representative of LBH.

These issues show that distinguishing different strata of BH is not easy, and these studies on Isa 40–66 make clear that distinguishing EBH from later BH can be problematic. In a review of "A Concise Lexicon of Late Biblical Hebrew" (Hurvitz et al. 2014) and an accompanying article, Rezetko and Naaijer (2016a and b) argue that many of the late linguistic features discussed in the Lexicon have limited value for describing LBH. There are only a few features in the Lexicon that occur in all the core LBH books, which means that many of the features described are rare and idiosyncratic features of BH. Instead of being characteristic of LBH, it is better to say that these features are characteristic of one book or a few books, and often the early alternative is used in the same late books as well. Also, the features described in the Lexicon are rare in the books that are supposed to be written in TBH. For instance, none of the features described in the Lexicon can be found in the book of Haggai. If it is supposed that the selection of late items in the Lexicon is representative of LBH, it is doubtful that there is a systematic difference between EBH and LBH and that biblical texts can be dated on the basis of language.

Another problem in most studies on linguistic variation on BH that we address is that there is continuity between EBH and LBH. Not only are the late features in LBH rare and idiosyncratic, but also in LBH there is a continuity of the use of

31 At least, according to Paul they occur in late literature, not according to Rooker, but that makes no difference for the argument here.

the early alternatives. Instead of being replaced by late alternatives, both early and late alternatives occur side by side in all the LBH books. This is a relevant issue for linguistics dating of biblical texts, because in descriptions of the method it is often said that early alternatives are replaced by late ones, but for most lexemes in the Lexicon, it is not so clear that this is really the case.

2.8. Conclusions

In the early 19th century, Gesenius proposed that one can distinguish between an early and a late variant of BH. Later scholars built further on this idea by distinguishing between ABH, EBH, and LBH. The disvovery of the DSS has added a whole new corpus to the relatively small corpus of Ancient Hebrew, and scholars have been able to show various links with different phases of BH, and within the phases of BH and QH variation has been discovered, for which a variety of explanations has been proposed. Also, other explanations were postulated for the linguistic variation in BH, such as regional and social variation. Based on the distinctive characteristics of mainly LBH, Hurvitz developed a method to date texts of unknown date on the basis of linguistic characteristics, which has been influential since the early 1970s. Since the late 20th / early 21st century linguistic dating has become a hotly debated topic. For Hurvitz, dating on the basis of linguistic characteristics is more objective than other approaches, but some scholars contest this assumption, and argue that language is only one criterion for dating texts, next to theological, historical and literary criteria. Other scholars argue that linguistic dating is invalid by criticizing the method itself.

Over the course of the past 200 years, it has become clear that the linguistic variation in BH is conditioned by numerous factors. At the same time, there is disagreement about what exactly we can know about the language on the basis of the extant evidence. This debate forms the background of this research. It is important to make a step forward, and process as much Hebrew data as possible, while taking into account that the linguistic variation in BH may have linguistic, geographic, social, historical, or literary backgrounds. It is not possible to deal with all these factors in one project, but in the Syntactic Variation project we have decided to choose four main variables as explanatory variables for linguistic variation. These will be discussed in the next chapter.

CHAPTER 3

Text, data, and scientific framework

3.1. The data

The dataset that is used throughout this research is an electronic edition of the Hebrew Bible, based on the ETCBC database. This electronic edition is called *Biblia Hebraica Stuttgartensia Amstelodamensis* or BHSA.[1] The ETCBC database contains the complete text of the fourth edition of the Biblia Hebraica Stuttgartensia and it is encoded completely on the levels of words, phrases, clauses, and text hierarchy. Next to the Hebrew Bible, the ETCBC database contains a number of DSS (1QM, 1QH[a] and 1QS), Rabbinic texts (Pirqe Avot and Shirata[2]) and Hebrew inscriptions.[3] These texts have been added to the ETCBC database by the project members as part of the Syntactic Variation project.[4]

The ETCBC database has been developed over a period of about forty years, and it is still in development. In the 1970s, Eep Talstra founded the "*Werkgroep Informatica Vrije Universiteit*" (WIVU) and started encoding the biblical books on the various linguistic levels. The history of the WIVU and its contribution to the study of the Hebrew Bible have been described by Oosting (2016), the data creation process has been described by Kingham (2017), and the data-model has been described by Talstra (2000), Talstra and Sikkel (2000), and Verheij (1994). Other references related to the data and their use can be found on the BHSA website.[5]

The tool that has been used throughout this research to extract data from the BHSA is Text-Fabric[6] (Roorda 2018). Text-Fabric is a Python package, developed by Dirk Roorda, with which annotated textual data can be preprocessed and analyzed

1 See https://pure.knaw.nl/portal/en/datasets/biblia-hebraica-stuttgartensia-amstelodamensis(eb177015-7e9e-4afe-8e49-1941cdab2a5f).html and https://github.com/ETCBC/bhsa.

2 This is the chapter "Shirata" from the Mekilta de-Rabbi Ishmael.

3 These are the Siloam inscription, 2 silver amulets from Ketef Hinnom, Lachish 3, 4, 5 and 6, the letter of complaint from Meṣad Hashavyahu, Kuntillet Ajrud 18.1, 19.1 and 20, the book of Balaam, son of Beor, and the Moabite Mesha stele.

4 As well as by Femke Siebesma in her project on verbal valence in the DSS.

5 https://etcbc.github.io/bhsa/references.

6 https://github.com/annotation/text-fabric.

further. It is the successor of an earlier system, LAF-Fabric.[7] An important difference between LAF-Fabric and Text-Fabric is the strongly simplified data format of Text-Fabric. It contains the text of the Hebrew Bible[8] in a simple, columnar text format. Additionally, it contains a complete grammatical analysis of the MT in separate text files, also in columnar format. In Text-Fabric, the data are modeled as a graph, in which the nodes, linguistic objects, are connected with each other by edges. In this way, words are connected to phrases, phrases are connected to clauses, and so on, up to the level of the biblical books. With this structure it is easy to find information about the clause in which a specific word occurs, and vice versa. With Text-Fabric it is possible to extract data from the BHSA and to save them in any desired format. The extrabiblical texts are also available in Text-Fabric format.[9] The main advantages of using Text-Fabric are its flexibility and its open source nature, making it possible for everyone to participate in the Text-Fabric project, and last but not least, it is completely free. Since the introduction of Text-Fabric a number of other text corpora have been added to the Text-Fabric ecosystem, such as texts written in cuneiform script[10] and the Qur'an.[11] These other corpora can be downloaded and processed separately.

There are two ways with which data can be accessed with Text-Fabric. The first approach is using Text-Fabric Search. Search is a template style query language.[12] Text-Fabric Search works in the browser. The second way of accessing the data is what I call the pure Python approach, which is the preferred option in this research.

As state above, texts are organized as objects in Text-Fabric. These objects have object-specific features and these features have a range of different values. The objects in the database are book, chapter, verse, half-verse, sentence, sentence-atom, clause, clause-atom, phrase, phrase-atom, subphrase, word, and morpheme. Most of these objects have the conventional linguistic meaning, but the atom objects need a bit of

7 LAF stands for Linguistic Annotation Framework and is essentialy an XML-based format to store annotated text. For LAF-Fabric, see https://github.com/Dans-labs/laf-fabric. LAF-Fabric is a conversion of the ETCBC database from Emdros. Emdros is a database engine for text data, which was developed by Ulrik Sandborg-Petersen. See https://emdros.org.

8 As an example of the data, see the file containing the lexemes in ETCBC transcription: https://raw.githubusercontent.com/ETCBC/bhsa/master/tf/2017/lex.tf.

9 https://github.com/ETCBC/extrabiblical; note that the format of the extrabiblical data differs slightly from that of the BHS data.

10 https://github.com/Nino-cunei/uruk.

11 https://github.com/q-ran.

12 https://annotation.github.io/text-fabric/Use/Search.

explanation. The difference between clauses and clause-atoms is relevant in cases in which a clause is embedded in another clause. An example can be found in Gen 9:18:

> ויהיו בני־נח היצאים מן־התבה שם וחם ויפת
> The sons of Noah, who went out of the ark, were Shem, Ham, and Japheth.

This sentence consists of two clauses. There is a היה clause, in which the attributive clause היצאים מן־התבה is embedded. In the ETCBC database, the main clause consists of two clause-atoms, the first one, ויהיו בני־נח, comes before the embedded clause, and the second one, שם וחם ויפת, comes after the embedded clause. So, a clause consists of more than one clause-atom if it has another clause embedded in it. To illustrate it a bit more, I have made a query that searches for all clauses in the MT that consist of at least two clause atoms.[13]

The datasets used in chapter 4 are made with Text-Fabric and stored in structured datasets in CSV format. Since this research is about clause structure, each row of the datasets contains the information about one clause, and each column contains information about a distinct feature. Sometimes the features are extracted directly from Text-Fabric, in other cases new features are derived from existing features. For instance, in the dataset containing two membered verbless clauses and היה clauses,[14] there is a column s_p_order, with the values s_p and p_s indicating the order of subject and predicate complement. This feature is not a clause feature of the ETCBC database, but is derived from the order of the phrases and the phrase feature "function".

The Text-Fabric scripts are made with Jupyter Notebook, and they can be found on my GitHub page, as can the resulting datasets and the scripts written in Python and R for all the calculations and visualizations in this research.[15]

13 https://shebanq.ancient-data.org/hebrew/query?id=1645. This query is written in MQL, the Mini Query Language, which is a query language developed by Constantijn Sikkel at the ETCBC for querying in the ETCBC-database. If the results are not visible, click on "4b" on the left side of the page.

14 https://github.com/MartijnNaaijer/phdthesis/tree/master/Ch4_Expressions_of_to_be/hyh_verbless/hyh_nom_bib.csv.

15 https://github.com/MartijnNaaijer/phdthesis.

3.2. The quantitative approach

The subtitle of this thesis is "a quantitative approach", which refers to a combination of statistical and machine learning techniques. In Biblical Studies, it is not usual to use statistical techniques to interpret linguistic variation in BH. This may be influenced by the old statement of Driver (1882: 203):

> Giesebrecht's facts are (with a few exceptions) correct: the use made of them is not sufficiently discriminating. The tabular synopsis is plausible and impressive: as the eye glances over it, the inferences which it is intended to carry home seem clear and unanswerable. The same may be said of the figures occurring so frequently in the later parts of the article. But both labour under a radical defect: they number words instead of weighing them; and when individual cases are examined, some cause which cannot be tabulated may appear for the presence or absence of a given word in a particular writing. In other words, the ostensible cause, apparent from the table or the enumeration, may not be the real cause which led to the employment of the word or phrase.[16]

It seems that its central emphasized sentence, "they number words instead of weighing them", has become an argument for avoiding quantitative research in general. At least, it has not had a stimulating effect on the use of quantitative methods. The problem, however, is that without "numbering words", it is possible to argue in any direction one wants, depending on one's presuppositions and choice of features. Comprehensive quantitative interpretations are a basic requirement of any balanced interpretation of linguistic variation in BH (and, for that matter, in studies of other languages), especially because BH is fairly homogeneous and relatively little is known about the historical background of the language and the transmission of the biblical texts.

In this research, I want to avoid this fear of statistics, and give it its rightful place in the study of linguistic variation in BH. Statistics has various goals, four of which I describe here.

16 Quoted by, among others, Hurvitz (2000a: 180); Rezetko (2003: 224 n. 17), and Hornkohl (2014: 41).

1. Quantifying uncertainty. In studies on diachronic variation in BH, generally two alternatives of a variable are given, one of which is said to be characteristic of LBH. Suppose two variables are considered, and one of them occurs four times and the other 100 times in the core LBH books. Of the first variable the late variant can be found three times in the LBH books and of the second one the late variant occurs 75 times in the LBH books. In this situation, the late variants of both variables have a prevalence of 75 % in the LBH books. In the case of the second variable, this percentage is more reliable, because the variable has a higher attestation. But how much more reliable is it? In classical statistics, this uncertainty is described with a confidence interval. The confidence interval is an interval around the mean value, in this case 0.75, which describes how uncertain the value of the mean is. In the case of a sample of four observations the confidence interval is wider than that of a sample of 100 observations. More information about confidence intervals can be found in Appendix A.

2. Making inferences. We only have a few biblical texts from the Second Temple period. We can assume that this is a small subset of all the texts from that period that have ever existed. With this subset or sample of texts it could be possible to say something about post-exilic Hebrew in general. Infering general properties about a whole population on the basis of a subset is done often in statistics because the population is too big to measure all individuals. In the case of Ancient Hebrew texts, we simply do not have more texts that have been transmitted, but the goal of inferential statistics is the same: finding out properties about a population on the basis of a subset of that population.

3. Testing hypotheses. Suppose someone wants to find out whether EBH and LBH are different with respect to some feature. Then a so-called null hypothesis (or H_0) is formulated, which says that there is no difference: EBH = LBH. With the help of a statistical test one can find out whether the null hypothesis should be rejected or not.

4. Making predictions. Machine learning generally deals with making predictions, and this is also common practice in Biblical Studies in the subfield of linguistic dating. In the case of a classification problem, one wants to predict the class of something. Linguistic dating is a typical example of a classification problem. In classification problems, one assigns a text to some target category as accurately as possible. In the case of linguistic dating of biblical texts, there are generally two main categories, EBH and LBH, and one wants to classify a book of unknown date as either EBH or LBH, based on a number of relevant features. Recent developments in machine learning have strongly increased the range of problems which can be treated using predictive techniques. In chapters 4 and 6, a number of prediction techniques are used. See also appendices A, D, E, and F.

Early explorers of quantitative methods in the field of BH are Radday and Forbes,[17] and in recent years there has been an increase of others who have used an approach that is more and more influenced by "numbering words".[18]

3.3. Open Science

In recent years, the scientific community has seen the emergence of a movement known as "Open Science",[19] to which this thesis wants to make a contribution. This movement is based on six principles that are all related to making the process of research, publishing, and education open to everyone:

1. Open methodology
2. Open source
3. Open data
4. Open access
5. Open peer review
6. Open educational resources

This movement began as the initiative of individual researchers, but it has reached a political level, where it is recognized as an important way forward in the scientific process.[20] Most of these aspects of Open Science form the basis of Text-Fabric and this research, and they are intended to solve important problems in current scientific practice.

In traditional studies on linguistic variation in BH, often one presents results without giving the underlying data. For instance, if a certain relevant word or syntactic structure occurs sixty times in the book of Genesis, the only thing that is reported is this number, without giving the complete list of attestations or the way those instances are collected. It is hard for the readers to check the data and to perform his or her own analysis of them. It is understandable that the full data were

17 See for instance the chapters 8–10 of Freedman, Forbes, and Andersen, eds. 1992.

18 Jacobs (2018), Rezetko and Young (2014) and some of the chapters in Miller-Naudé and Zevit, eds. 2012. Chapter 5 of Young, Rezetko, and Ehrensvärd (2008) created some discussion about the proper sample size if one wants to have a representative sample. This is an interesting issue that needs to be discussed more often in Biblical Studies, because the answer may not be so straightforward.

19 http://openscienceasap.org/open-science.

20 https://ec.europa.eu/research/openscience/index.cfm.

not given in printed journals and books, because this would have led to enormous appendices, but with the advent of the Internet it is easy to store large amounts of data.[21]

3.4. Linguistic framework

This research is based on a form-to-function linguistic approach, which is strongly related to the origins and development of the ETCBC database by Eep Talstra. In Talstra's opinion, little is known about Hebrew syntax (Van der Merwe 1997: 13), so it is best to start with a description of the formal characteristics of a text. He uses this approach not just on the levels of words and morphology, but also on the syntactic levels of phrases, clauses, and relationships between clauses. Based on this, language can be analyzed with the use of computer programs (Talstra 1991), which as a pioneer he advocated for since the 1970s. A focus on the formal characteristics is, of course, the most obvious choice for an algorithmic analysis, because a computer can easily read a text as a sequence of strings and analyze it as such.

A biblical scholar working earlier from this perspective is Hoftijzer, who advocated a formal approach to the study of verbless clauses in his review of Andersen's study of verbless clauses in the Pentateuch (Andersen 1970; Hoftijzer 1973). In this work, Hoftijzer opts for describing the verbless clause in formal terms. Instead of describing it in terms of phrase functions (a juxtaposition of subject and predicate complement), it is better to stick to a lower level description of phrase types (for instance, NP-PP).

In the present research, the formal approach to syntax has the advantage that it avoids functional subtleties (Van Peursen 2007: 142) that are difficult to quantify. It must be stressed that there are various new developments in NLP, which make it possible to capture semantic information algorithmically. These techniques are based on the idea that the meaning of a word depends on the words surrounding it; they are called embedding techniques, such as word2vec (Mikolov et al., 2013).

I have chosen to study the linguistic variation in BH within the framework of corpus linguistics. In general, corpus linguistics deals with data, in our case textual data, that is on such a scale that it is impossible to inspect and investigate everything manually. The corpus under investigation consists of a set of texts in which one searches for (linguistic) patterns and pattern variation. The variation in patterns may

21 For instance, on GitHub: https://github.com, or in open repositories funded by governmental organizations, for instance: https://dans.knaw.nl.

be an indication of the chronological development of a language or of geographical or some other type of variation.

Ideally, a corpus is balanced and representative: it is balanced in the sense that one searches for sources of diverse origin; it is representative in the sense that the proportions of the samples from certain sources in the corpus are more or less representative for the proportions in "real life". In the case of newspapers, if two different newspapers have more or less the same number of articles every day, more or less equal numbers of samples should be chosen from these newspapers (McEnery and Hardy 2012: 8–9).

In practice, it is difficult to create a balanced corpus, because even if there are many sources available, it is possible that there is substantial variation within a subcorpus of closely related texts (Gries, 2006). It is clear that the corpus of Ancient Hebrew is not representative, nor balanced. The whole corpus is relatively small,[22] it contains mainly texts originating in Judah[23] in priestly circles, and it is absolutely unclear how representative different texts are for some kind of language variety. The corpus of Ancient Hebrew is an opportunistic corpus (McEnery and Hardy 2012: 110): we study the texts that we study, because there are no other texts available.

3.4.1. *The clause as a sequence phenomenon*

In this research, clauses are treated as sequential structures, especially in chapter 6. Often in linguistic research, clauses are analysed using hierarchical structures (trees). I do not deny that hierarchical analysis can be valuable in the analysis of language, but the sequential approach offers some advantages. In the first place, a sequence is a simpler structure than a tree. If a sequential model and a tree model produce equal results, applying Occam's Razor leads to a preference for the simpler model. Also, recent developments in machine learning make it possible to process sequential data in a much more advanced way than just analysing at word level or n-grams (Franka and Christiansen 2018: 1217). For instance, with an LSTM model, such as the one used in chapter 6 of this research, it is possible to model complex long-term dependencies in sequences. It would be interesting to compare sequence models and tree-based models in BH, but such a comparison is beyond the scope of this research. One of the features in this research is "mother" (see section 4.2.4 for explanation), which is

22 By far the largest sub-corpus of Ancient Hebrew is formed by the Hebrew Bible, which contains about 426,000 words.

23 Rendsburg estimates that about 80% of the Hebrew Bible is written in the Judean dialect.

actually based on a hierarchical model of language on the level of larger text units. Also, the main variable, main and dependent clauses, is partly based on the ETCBC database feature CARC (clause atom relation code). So, for the embedding of a clause the text hierarchy of the ETCBC database is used, the structure of the clause itself is modelled as a sequence.

3.5. Independent variables in the research

The three subprojects of the Syntactic Variation project deal with different issues and use different approaches. However, there is a constant factor uniting them, which is that all seek to interpret syntactic variation in the light of four main independent variables.

3.5.1. *Genre*

We distinguish between three main genres: prose, poetry and prophecy. The genres have been assigned to complete books, based on how the genres are used generally in exegetical literature.

In the order of books of the Hebrew Bible, the following main division is used:

Prose: Genesis, Exodus, Leviticus, Numbers, Deuteronomy, Joshua, Judges, 1 Samuel, 2 Samuel, 1 Kings, 2 Kings, Jonah, Ruth, Esther, Daniel, Ezra, Nehemiah, 1 Chronicles, 2 Chronicles.
Poetry: Song of Songs, Proverbs, Qoheleth, Lamentations, Psalms, Job.
Prophecy: Isaiah, Jeremiah, Ezekiel, Hosea, Joel, Amos, Obadiah, Micah, Nahum, Habakkuk, Zephaniah, Haggai, Zechariah, Malachi.

We have refined this division by giving different values than these defaults to specific texts. Examples are Gen 49, Exod 15, Judg 5, 1 Sam 2:1–10 that have been given the value poetry.[24]

In the datasets made with Text-Fabric, this variable is called "genre", having the values "prose", "poetry", and "prophecy". The distinction of three genres is rather crude, but it works well as a point of departure. Separating the data in more genres and

24 The complete list can be found in a file that was made by Dirk Bakker: https://github.com/MartijnNaaijer/phdthesis/blob/master/Various/subgenres_synvar.xls.

subgenres would result in many debatable splits, which would also lead to problems in the statistical analysis.

3.5.2. *Discourse type*

The second main independent variable is the discourse type of a clause. In the ETCBC database, four discourse types can be distinguished. These are "N" (narrative), "Q" (quoted speech), "D" (discourse), and "?" (not determined).[25] A chapter starts with ? if there is no clear indication of the communication situation. The discourse type of a clause can change to N if there is a clause with a wayyiqtol form. An example can be found in Gen 4:1.

והאדם ידע את־חוה אשׁתו ?	And Adam knew Eve, his wife.
N ותהר	And she conceived.
N ותלד את־קין	And she bore Cain.

The first clause of Gen 4:1 contains no clear indication of the communication situation, but the second clause contains a wayyiqtol, so the value of the discourse feature becomes N. The third clause also contains a wayyiqtol, so the value stays N. The situation changes in the following clauses of Gen 4:1.

N ותאמר	And she said:
NQ קניתי אישׁ את־יהוה	I produced a man with YHWH.

The wayyiqtol form in the clause ותאמר is a verb of speaking and indicates that a discourse change follows in the next clause, in which Eve speaks. This quoted speech is indicated with the letter Q, and because it is embedded in narrative the value is NQ. These embeddings can become deeper, NQQ indicates that there is quoted speech within quoted speech within narrative.

The last discourse value is D, assigned when the narrator addresses the reader directly with some background information. A formal characteristic of D is that a D clause has a yiqtol form within a narrative environment. An example can be found in Gen 2:23–25.

25 These values are assigned automatically using an algorithm on the basis of linguistic characteristics present.

N ויאמר האדם	Then the man said:
NQ זאת הפעם עצם מעצמי ובשׂר מבשׂרי	This at last is bone of my bones and flesh of my flesh.
NQ לזאת יקרא אשׁה	This one is called woman.
NQ כי מאישׁ לקחה־זאת	For from man this one was taken.
ND על־כן יעזב־אישׁ את־אביו ואת־אמו	Therefore a man leaves his father and his mother.
ND ודבק באשׁתו	And clings to his wife.
ND והיו לבשׂר אחד	And they become one flesh.
N ויהיו שׁניהם ערומים האדם ואשׁתו	And they were both naked, the man and his wife.
ND ולא יתבשׁשׁו	And they were not ashamed.

In this section, the foreground comprises the two clauses with a wayyiqtol, these clauses are assigned the value N. There is also embedded quoted speech (NQ) and there are some parts where information is addressed directly to the reader (ND), introduced by a yiqtol.[26]

The embedding of discourse types can lead to deeply embedded encodings, that are not really relevant for the present research. Therefore, only the last letter of the feature value will be used in the analysis. For example, if the feature value is NQ, in the analysis it is treated as Q. If the value is N, then it is treated as N.

In the datasets made with Text-Fabric this variable is called "txt_type", having the values "N", "Q", "D" and "?".

3.5.3. *Language phase*

The third main variable is the supposed language phase. It is debatable how to encode this, but we have chosen to use an encoding that is relatively conventional. The traditional categories of EBH and LBH are used to test if this distinction can be observed in the syntactic features studied.

The books of the Pentateuch and the Former Prophets (Genesis, Exodus, Leviticus, Numbers, Deuteronomy, Joshua, Judges, 1 Samuel, 2 Samuel, 1 Kings, 2 Kings) are classified as EBH and the core late narrative books (Esther, Daniel, Ezra, Nehemiah, 1 Chronicles, 2 Chronicles) as LBH. The other biblical books get the value "other". The

26 This explanation of the various discourse types in the ETCBC database is a somewhat simplified version of how it works.

extra-biblical books have the values "epigraphic" and "qumranic" for, respectively, the pre-exilic inscriptions and 1QS, 1QH^{a}, and 1QM, and Pirqe Avot and Shirata have the value "rabbinic". In the datasets that are extracted using Text-Fabric this variable is called ebh_lbh having the values "ebh", "lbh", "other", "qumranic", "epigraphic" and "rabbinic".

3.5.4. *Main and subordinate clauses*

The fourth independent variable is that of main and subordinate clauses. In our research, we want to base the distinction between main and subordinate clauses as much as possible on formal criteria, and also, we want to do justice to the various kinds of subordinate clauses that can be distinguished. The distinction used here is developed by Marianne Kaajan, and described in detail in Kaajan, "Main and subordinate clauses in Biblical Hebrew" (in preparation). A distinction is made between main and subordinate clauses, and a further distinction is made between three types of subordinate clauses: argument clauses, relative clauses, and adverbial clauses.[27]

An argument clause is a constituent of another clause. This role is fulfilled generally by a noun phrase. Examples are:

> Gen 2:18 לא־טוב היות האדם לבדו
> It is not good that the man should be alone

Here the infinitive clause היות האדם לבדו is the subject of the main clause.

> Exod 7:5 וידעו מצרים כי־אני יהוה
> The Egyptians shall know that I am YHWH.

The clause כי־אני יהוה is the object of the main clause.

Next to object and subject clauses, the ETCBC distinguishes the following functions an argument clause can have in another clause:[28]

---

27 The distinction between main and the three types of subordinate clauses was implemented using Text-Fabric in this script: https://github.com/MartijnNaaijer/phdthesis/blob/master/Various/main_subordinate_clauses.ipynb.

28 The argument clauses are underlined.

Predicate Complement clause
Exod 3:14 אהיה אשר אהיה
I am who I am.

In this short sentence, אשר אהיה functions as Predicate Complement of the main clause.

Complement clause
Gen 15:5 אם־תוכל לספר אתם
If you are able to count them.

In this sentence, לספר אתם is a complement of the verb in the main clause. It cannot be removed without making the remaining sentence ungrammatical.

The second class of subordinate clauses that we distinguish are attributive clauses, which modify a noun phrase. These are frequently introduced by the so-called "relative pronoun" אשר. An example:

Judg 4:22 ואראך את־האיש אשר־אתה מבקש
And I will show you the man whom you are seeking.

The third class consists of adverbial clauses. These clauses add optional information to the main clause. Some examples (the underlined clauses are adverbial clauses):

Num 22:3 ויגר מואב מפני העם מאד כי רב־הוא
Moab was in great dread of the people, because they were so numerous.

Ruth 3:9 ופרשת כנפך על־אמתך כי גאל אתה
Spread your cloak over your servant, for you are next-of-kin.

In the datasets made with Text-Fabric, this variable is called "main_sub" and has the following values: "Main" (main clause), "SubArg" (subordinate, argument clause), "SubAdv" (subordinate, adverbial clause), "SubMod" (subordinate, attributive clause).

The difference between adverbial clauses and argument clauses is that the latter type is a necessary part of a clause whereas the adverbial clauses can be omitted without making a clause ungrammatical. There is actually a wide variety of subordinate clauses which fall under the category of adverbial clauses in our research, and linguists differ in how they describe the adverbial clause.

TABLE 3.1 *Categories of main and subordinate clauses and degrees of finiteness*

Form → *Function ↓*	Finite	Infinitival	Participial	Nominal
Main	main-Fin	main-Inf	main-Ptc	Main-Nom
Subordinate - Argument of verb	suboArg-Fin	suboArg-Inf	suboArg-Ptc	suboArg-Nom
Subordinate - Modifier in NP	suboMod-Fin	suboMod-Inf	suboMod-Ptc	suboMod-Nom
Subordinate - Adverbial	suboAdv-Fin	suboAdv-Inf	suboAdv-Ptc	suboAdv-Nom
Extra-clausal	excl-excl			

Besides the distinction between main and three types of subordination, the Syntactic Variation project distinguishes different degrees of finiteness of a clause. A distinction is made between finite, infinitival, participial and nominal clauses, and these respective categories are considered to be in decreasing order of finiteness.[29] The distinction between main and three types of subordinate clauses and between four levels of finiteness leads to sixteen different categories of clauses, which are shown in table 3.1.[30]

There is a seventeenth category of extra-clausal elements. These are elements that are categorized as clauses in the ETCBC database, because in the database all words need to be part of a clause, but they lack predication. Examples of such are macrosyntactic signs and vocatives.

29 For a more elaborate description, see Kaajan, "Main and subordinate clauses in Biblical Hebrew" (in preparation).

30 The table was made by Marianne Kaajan, see Kaajan, "Main and subordinate clauses in Biblical Hebrew" (in preparation).

Some examples:

> Gen 4:23 (vocative) ויאמר למך לנשׁיו עדה וצלה שׁמען קולי נשׁי למך
> Lamech said to his wives: "Adah and Zillah, hear my voice; you wives of Lamech"

> Gen 3:22 (macrosyntactic sign) ועתה פן־ישׁלח ידו
> And now, he might reach out his hand.

In my research, the degrees of finiteness are not taken into consideration, because variation in finiteness is implied in the different classes of the dependent variable in chapter 4.

3.6. Syntactic variation and experimental approach

In recent years, some scholars have opted for a variationist approach in studying linguistic variation in BH (Kim 2013; Rezetko and Young 2014). This approach studies how something can be said in different ways. A well-known example is the word "kingdom", which can be said in at least two ways in BH (ממלכה and מלכות). Variationist analysis tries to describe and explain the distribution of these different expressions and tries to link the linguistic phenomena to other linguistic and extra-linguistic phenomena (Rezetko and Young 2014: 215).

Rezetko and Young (2014) mention the following drawbacks of studying syntax in the context of variationist analysis:

1. Some syntactic units may require large quantities of writings to obtain enough tokens to find patterns of usage (frequency requirement).
2. Syntax is abstract and sometimes it may be difficult to define the variable context (or semantic equivalence) (comparability requirement).
3. Syntax is sensitive to genre differences and so controlling for genre is crucial when comparing different constructions (genre requirement).
4. Syntactic change is gradual or, conversely, syntax is relatively stable, so that writings covering a long period of time are usually needed in order to uncover development (time requirement).
5. Sometimes other disciplines—Hudson gives several examples related to psycholinguistics and discourse analysis—may offer better explanations for the differences

> between two "synonymous" syntactic constructions (Rezetko and Young 2014: 229, quoting Hudson 1996: 172).

According to what has been said above, one of the main problems of studying variation is that the variants should be more or less semantically equivalent. For this reason, Labov found phonological variables the most useful. Problems arise when studying the lexicon. For many lexical features it is not entirely clear to what extent the alternatives are semantically similar. For instance, Bergey gives כתר (crown) as a late alternative of the EBH nouns נזר and עטרה (Bergey 1983: 98–99). What does the word mean? From the context it is clear that כתר is something someone can put on his or her head, but does this imply the same nuances as when the author would have used one of the early alternatives?

Similar issues arise in the study of syntax. It is not always clear whether there are semantic alternatives of a certain variable. In chapter 4, syntactic alternatives of expressions of "to be" are studied. These expressions are more or less clear syntactic alternatives. There may be semantic variation between these alternatives, but detecting this on the basis of the formal characteristics of the environment of the clauses can be one of the outcomes of the research. In chapter 6, syntactic variation is approached by studying all clauses from two subcorpora (EBH and LBH), and specific characteristics of each of these subcorpora are extracted automatically by the algorithm. This approach works around the problem of semantic alternatives.

Traditional variationist sociolinguistics—in which linguistic variation is linked to social variation empirically—generally uses techniques like variable rules analysis, which is basically multivariate logistic regression. It can be found in packages like Goldvarb.[31] In this research, I choose more modern and flexible tools.

3.7. Conclusions

In this chapter, I have discussed the various aspects of the framework of this research. The most important and innovative of these are Open Science, and the quantitative and multivariate approach of the work. The objective and approach of this research stands in line with conventional historical linguistic work in the sense that it is empirical and descriptive.

31 http://individual.utoronto.ca/tagliamonte/goldvarb.html.

The process and results of scientific research should be transparent and everyone should be able to check how the research was done. In this research, I use the Text-Fabric software and the BHSA data. This combination is ideally suited for sharing data, for instance, on GitHub.

The quantitative approach is the most important methodological step of this research. Instead of focusing on rare features, this research has been designed around frequently occurring structures. These are analyzed using various packages of the Python and R programming languages and patterns in the data are clarified by visualizing them. This approach of studying linguistic variation is an empirical and quantitative enterprise, and therefore, if it is treated mainly qualitatively, only half of the work is done. "Quantitative" is still a general designation, and depending on the specific problem, one can choose various approaches.

CHAPTER 4

Alternative expressions for "to be"

4.1. Introduction

This chapter deals with the BH expressions of "to be", for which BH has five common alternatives. One of these uses a form of the verb היה, the others use the particles יש ("there is") and אין ("there is not"), or take the form of a bipartite or tripartite verbless clause. These various ways of expressing "to be" can be found in diverse structures, environments, and semantic nuances. The main research question in this chapter is:

> How is the variation in the use of the various expressions of "to be" in BH conditioned?

Of the four types of clauses that are discussed in this section, bipartite verbless clauses are by far the most frequent, occurring in a wide variety of constructions. Therefore, they are seen as the default type. In each of the following three subsections, two syntactically related types of clauses will be discussed. In each subsection, bipartite verbless clauses are contrasted with a less frequent syntactic alternative.[1] Section 4.2 deals with היה clauses and bipartite verbless clauses, in section 4.3, verbless clauses with and without יש are investigated, and in section 4.4, bipartite and tripartite verbless clauses will be analyzed. In each of these sections, only complete clauses containing both a subject and a predicate complement are studied. For each of the three analyses, a distinct dataset is prepared. Each dataset contains information about bipartite verbless clauses and one of the other clause types. The bipartite clauses are selected in each analysis in such a way, that the variation in their structure is similar to the variation in the structure of the alternative. Clauses with היה occur in a wide variety of structures, so bipartite clauses are selected with a similar variety. Clauses with יש are less varied. Therefore, the bipartite clauses without יש with which they are compared, are selected in such a way that they have similar structures as the clauses with יש.

1 The frequencies of the various clause types can be found in table 4.2 (section 4.2) for clauses with and without היה, table 4.4 (section 4.3) for clauses with and without יש, and table 4.5 (section 4.4) for bipartite and tripartite clauses.

4.2. היה clauses and bipartite verbless clauses

4.2.1. *Problem and research question*

The two most frequent realizations of "to be" in BH are clauses with the verb היה and verbless clauses. From past research it has become clear that the verb היה can have various functions in the clause. It is related to tense, aspect, and mood (TAM), it is often only a placeholder verb without meaning, but in some cases, היה should be translated by an English verb like "to become". Generally, היה clauses and bipartite verbless clauses have been studied in relative isolation. When היה clauses are discussed, the clause is analyzed and syntactic and semantic details are studied. Such an analysis of the היה clauses reveals much about the function of היה in the clause, but it does not necessarily bring to light the difference between היה and verbless clauses. The same applies to studies of verbless clauses: studying structural variation between verbless clauses does not necessarily show how a verbless clause with a certain structure relates to a clause with a similar structure in which the verb היה is present.

On the syntactic level, there do not seem to be strong differences between היה clauses and verbless clauses. Both can occur with noun phrases, demonstrative pronouns, and personal pronouns as subject, and the predicate complement can consist in both clause types of a noun phrase, adjective phrase, or prepositional phrase.

Both clause types occur in all of the biblical books, so there does not seem to be a clear linguistic development in the use of one or the other. However, although there is no sharp distinction in the structure and environment of verbless clauses and היה clauses, no systematic research has been done on the influence of the syntactic structure and the environment of verbless clauses and היה clauses. For instance, both clause types occur with a subject NP and a predicate complement PP, but could it be that such a structure has a preference for one of the two alternative realizations? Similarly, when there are more constituents in the clause, does that influence the preference for היה or not? How does the environment of the clause influence the use of היה? The research question for this section is therefore:

> Which clause-internal and clause-external factors influence the choice to use היה or not?

The main focus of the research will be on the central variables of the Syntactic Variation project, distinguishing main and subordinate clauses, genre, language phase, and discourse type. A number of other relevant variables will also be taken into consideration.

4.2.2. *Review*

Expressions for "to be" in BH are often discussed in the context of the copula. This chapter could also have been called "The use of the copula in BH", but there are varying opinions as to whether the five expressions of "to be" in BH can be interpreted as copula. Therefore, the more general term "to be" is used.

According to Dixon (2010: 159), the copula clause is a third clause type next to transitive and intransitive clauses. The copula clause is characterized by a copula verb which has two core arguments, the copula subject and the copula complement. The possibility to have these two core arguments is even the defining feature of the copula verb (Dixon 2010: 160). Dixon gives two relations that are always covered by the copula: identity ("This man is a doctor") and attribution ("This man is clever"). In many languages, possession ("This book is John's"), benefaction ("This present is for John's birthday"), and location ("The apple tree is in the garden") are also relations covered by the copula clause (Dixon 2010: 159–160). There are languages that have a particle that means "there is", but this should not be considered a copula, according to Dixon, because it is an existential use of "be" (Dixon 2010: 160). An important difference between verbless clauses and copula clauses is that the former do not mark tense (Dixon 2010: 161). Copula clauses and verbless clauses are similar and, in many languages, e.g., Hebrew, both occur.

In languages that have both copula clauses and verbless clauses, the omission of the copula is often explained by pointing to the TAM characteristics that are carried by the morphology of the copula. In various languages, the verbless clause has present tense, while the past and future are marked forms. In these cases, the copula is used, for instance in Russian (Dixon 2010: 181). There are languages in which the copula is required under certain conditions, and for other languages there is a looser conditioning of its use (Dixon 2010: 181).

The use of היה in BH comes close to Dixon's description of the copula, although among linguists of Hebrew opinions vary on the precise interpretation of the use of the verb. According to most grammars of BH, the main difference between the verbless clause and the היה clause is that היה is added to mark the clause explicitly with TAM, as present in the morphology of היה. One can distinguish between views that see this as the only function of היה in this type of clause and views according to which היה has more functions/meanings than that.

A good example of the former position is found in Waltke and O'Connor's *An Introduction to Biblical Hebrew Syntax* (1990). In their opinion, in languages with an optional copula, like in Hebrew, Greek, or Latin, the explicit copula "is thus to mark

in the surface structure tense, mood, or aspect" (Waltke and O'Connor 1990: 72). Bartelmus (1982) dedicates a monograph completely to the verb היה. For Bartelmus, the verb היה is not a true verb, it has no participle[2] and it occurs in structures that closely resemble nominal clauses. It only adds TAM to these clauses: "*Wenn* היה *selbst keinen spezifischen lexikalisch-semantischen Wert hat, kan es in beliebigen Zusammenhängen—nicht nur zur Verzeitung von verblosen* NS*—als Träger der tempusrelevanten grammatischen Morpheme, d.h. als blosser Tempusmarker eingesetzt werden*" (Bartelmus, 1982: 114). This means that translating היה is problematic, because its presence is more an issue of syntax than of the lexicon. However, with the lack of better options one uses the German "sein" (Bartelmus 1982: 114), but one should remember that "*HYH hat keine eigenständige lexikalisch-semantische Bedeutung, sondern dient lediglisch zur 'Verzeitung' bzw. 'Modifizierung' von Sachverhalten, die bei Zeitbezug* GZ *mit verblosen* NS *ausgedrückt werden*" (Bartelmus 1982: 114). In the opinion of Bartelmus, it is not correct to speak of היה as a copula, because it is used in situations in which the German "sein" is never used, for instance, in the case of "Ingress", when in general "werden" is used, and also because this terminology brings the Indo-European terminology and underlying philosophy to a completely different situation (Bartelmus 1982: 96–97). JM follows Bartelmus in §154m. היה can specify the temporal sphere or express a volitive mood. In the latter case, the jussive יהי is used, although it can be omitted without losing its optative force (§163b),[3] and in various cases a clause without a form of היה can refer to the past or the present, depending on the environment, e.g. ויוסף בן־שלשים שנה, "And Joseph was thirty years old" (Blau 1993: 84). Schoors also follows Bartelmus, as does Niccacci (1990). Nicacci explicitly says that היה in a qatal form refers to the past, and in a yiqtol form, refers to the future. Schoors (2004) studies the meaning and function of היה in the book of Qoheleth. He argues, following Bartelmus, that in general היה has no meaning, but it "acts as a tense marker for a nominal clause" (Schoors 2004: 51). However, in certain cases in the book of Qoheleth, Schoors observes that היה does have semantic content, for instance, היה בטוב in *7*:14 means "to live well, to be in a good mood" and היה צדיק in Qoh *7*:16 means "to act rightly" (Schoors 2004: 51). The cases of

2 Bartelmus overlooks the participle in Exod *9*:3 הנה יד־יהוה הויה במקנך, "The hand of YHWH is on your cattle".

3 JM gives the examples of Judg *6*:23 שלום לך, "Peace be with you", Ruth 2:4 יהוה עמכם, "YHWH be with you", and the participal clauses Gen 3:14 ארור אתה, "Cursed are you", and Gen *9*:26 ברוך יהוה, "Blessed is YHWH".

the verb with the meaning "to happen" show that the author is a real philosopher (Schoors 2004: 59).

Sinclair adopts Waltke-O'Connor's description of the function of היה, but suggests some modifications (Sinclair 1999: 52). Sinclair recognizes that the verb היה, like the English verb "be", can function in many ways beyond a mere copula. These different functions result in English glosses like "happen, occur, fall upon, come, come to pass, become", etc. Rather than viewing these senses as "definitions" of the verb היה, Sinclair argues that we should understand them simply as translation glosses required for idiomatic English (Sinclair 1999: 53–54, see also GKC, § 154i, and JM, § 111i, for similar examples of semantic nuances of היה).

4.2.3. *Regression analysis*

In this section, the research problem is how the use or non-use of היה in bipartite clauses relates to a number of other variables, of which the most important are genre, language phase, discourse type, and whether or not a clause is a main or subordinate clause. Do these variables influence the use of היה? This type of problem is generally approached with regression analysis, which is also the way it will be dealt with here. By bringing all the predictors together in one multivariate model, one can get a better understanding of the relationship between the predictors and the output variable, than when the predictors are studied in isolation in relation to the output variable.

The regression technique that is used in this analysis is Generalized Additive Mixed Modeling (GAMM, see Hastie and Tibshirani 1990; Wood 2017) with a binary output. This technique is relatively new in the study of linguistics and is used mainly in experimental studies to investigate non-linear movement of the tongue (Wieling et al. 2016) or eyes (Nixon et al. 2016). This is done by using a smoother on continuous variables, whereby the assumption of a linear relationship between dependent and independent variables is relaxed. In the present research, this advantage of GAMMs is not relevant, because continuous variables are hardly used here, but another advantage of GAMMs is. In the dataset, a clause is sampled repeatedly from the individual books. Therefore, the clauses are not independent samples, because they are clustered in the books. This violates the independence assumption of a Generalized Linear Model (GLM). With a GAMM it is possible to account for repeated measurements in individuals (in experimental research these individuals often are the experimental subjects), and thus use the smoother in a numeric variable (in this case, the clause id) as random effect.

An additional problem is also dealt with by using a GAMM. In time series or sequence data, which literary texts typically are, one often encounters autocorrelation, for which a GAMM can be a solution (Baayen et al. 2016).[4]

With the mgcv package it is possible to use various smoothers. In this research, I use a factor smoother to account for individual variability between biblical books. The smoother functions as a random effect with random intercept and random slope. By doing this, p-values are prevented from being too low without sufficient justification.

For more information about regression analysis, see Appendix B.

4.2.4. *Variables*

The Text-Fabric scripts and the resulting datasets can be found on GitHub.[5]

The dependent variable

The dependent variable in this investigation is clause type (cl_type).[6] The variable has two possible values: "hyh" (the clause contains היה) and "nom" (the clause does not contain היה). All the clauses under investigation contain a subject and a predicate complement, and all the predicate complements are non-verbal, so participial clauses (with or without היה) are excluded from the analysis. Also excluded from the analysis are clauses containing יש and tripartite verbless clauses, because these will be analyzed in a later section.

The independent variables

The following independent variables are studied in this research. The presence or absence of היה in a clause is a clause level issue, but it may be influenced by factors at a higher level. Therefore, I distinguish between clause-internal and clause-external variables.

---

4 The data are analyzed using the function bam from the mgcv package. This function is similar to the function gam, but it is optimized for large datasets https://cran.r-project.org/web/packages/mgcv/mgcv.pdf.

5 In the folder https://github.com/MartijnNaaijer/phdthesis/blob/master/Ch4_Expressions_of_to_be/hyh_verbless. Text-Fabric notebooks: hyh_verbless_bib.ipynb and hyh_verbless_xbib.ipynb. Biblical data: hyh_nom_bib.csv. Extrabiblical data: hyh_nom_xbib.csv.

6 The variable name as it can be found in the dataset is italicized.

Clause internal variables

The following clause-internal variables can be found in the dataset (the name of the variable in the dataset on GitHub is between parentheses; not all the variables are included in the final model):

Clause length (*cl_len*) refers to the number of phrases a clause contains. In the היה clauses, the היה phrase is excluded in the count, because, by definition, these clauses have one phrase more than verbless clauses. The minimum clause length is obviously two, which is the length of a clause with only a subject and predicate complement.

Subject length (*subj_len*) is the number of words in the subject. Pronominal suffixes are counted as separate words.

Predicate complement length (*pc_len*) is the length of the predicate complement. As in the case of the subject, the length is counted in words, and again, pronominal suffixes are counted as words. A prepositional phrase like בַּמִּדְבָּר is counted as three words, preposition, article, and noun.

Subject-Predicate Complement order (*s_p_order*) is the order of the subject and predicate complement.

Subject type (*subj_type*) is the phrase type of the subject.

Subject definiteness (*subj_det*) is the definiteness of the subject. One can distinguish formal and semantic definiteness (also called determination, Baasten 2006: 39–40). In this research, it is approached from a formal perspective.

Predicate complement type (*pc_type*) is the phrase type of the predicate complement.

Predicate complement definiteness (*pc_det*) is the definiteness of the predicate complement.

Language is the language of the clause.

In the analysis, only Hebrew clauses are taken into account. Most clauses in the dataset are relatively short, consisting of a subject, a predicate complement, and in some cases the verb היה. However, the clause can contain other phrases. The datafile contains columns for all the different phrase functions that occur in the Hebrew Bible, besides subject, predicate complement, and predicate. These are binary variables (1 for present in a clause, 0 for absent).

Clause-external variables

The clause external variables are variables that are related to the environment of the clauses in the dataset. The main variables of the Syntactic Variation project are included in the dataset and the analysis described in section 3.5. Other variables that are taken into account are:

Book is the book in which a certain clause occurs. Book is considered a random variable in this mixed model, because the clauses do not exist as independent events, but are clustered in the book in which they occur. In the analysis, the books of Samuel, Kings, and Chronicles are each considered one book. This is done because these books were historically one book, and treating them as two books is actually more artificial than merging them together. The dataset has two columns for the bookname. The column book uses the names 1 and 2 Samuel, 1 and 2 Kings, and 1 and 2 Chronicles. The column book2 only uses the names Samuel, Kings, and Chronicles. The latter column is used in the analysis. The choice of using complete books in the analysis does not necessarily mean that they are considered to be literary unities. There are many ways in which the books can be split in redactional layers, and doing this for the whole Hebrew Bible would be a task which would lead to too many arbitrary splits. Therefore, the choice is to use complete books, a choice which has at least an empirical basis. The dataset also includes columns for chapter and verse.

Clause id (*cl_id*) is the clause identifier. It is the node number of the clause in Text-Fabric.

Mother (*mother*) represents the tense of the clause on which the clause under consideration is dependent syntactically. This idea of dependency of clauses on each other is implemented in the ETCBC database and is based on the ideas of the text hierarchy of Schneider (Schneider and Grether 1974). It was maintained by Kalkman in his PhD thesis (Kalkman 2015) that the meaning of Hebrew verbs depends strongly on the hierarchical relationships between clauses. Kalkman is mainly concerned with BH poetry and more in particular with the Psalms, but he suggests that these ideas are applicable to prose as well. If this is the case, it is well worth investigating whether text hierarchy may influence the use of היה.

Clause relation (*clause_rela*) is the clause relation. It is relevant for the exclusion of tripartite verbless clauses from this analysis (resumptive clauses following a casus pendens, in which the subject is a third person pronoun). It is also relevant for the decision whether a clause is a main or a subordinate clause.

TABLE 4.1 *Frequencies of clauses with and without* היה

	Epigraphic	MT	Qumran	Rabbinic
Clauses with היה	7	3554	122	169
Clauses without היה	391	83274	4768	3451

4.2.5. *Data preparation and variables*

The dataset consists of the היה clauses and verbless clauses containing a subject and a predicate complement in the Hebrew Bible, 1QS, 1QH[a], 1QM, Pirqe Avot, and Shirata, as available in Text-Fabric.[7]

Each row of the dataset contains information about one clause. In the dataset, the output variable cl_type (clause type) has two values. It has the value "hyh" if the clause contains the verb היה, and "nom" if the clause is a verbless clause.

4.2.6. *Data exploration*

Clauses with and without היה

Before moving to clauses with both subject and predicate complement, all the clauses in the MT with and without the verb היה are explored.[8] This gives a general impression of the frequency of the use of היה. The raw numbers for the subcorpora under consideration can be found in table 4.1 (see next page).

With a mosaic plot, one can get an intuitive impression of how the levels of categorical variables are associated with one another. This is done by comparing the expected frequencies with the observed frequencies of different combinations of the relevant variables. The figures 4.1–4.4 show the distribution of clauses with and without the verb היה, based on the four core variables of the Syntactic Variation project.

The surface of each rectangle represents the frequency of a specific clause type in each of the phases of Classical Hebrew. This makes it possible to compare the

7 Scripts and datasets can be found here: https://github.com/MartijnNaaijer/phdthesis/tree/master/Ch4_Expressions_of_to_be/hyh_verbless.

8 The clauses without היה are both verbal and non-verbal clauses.

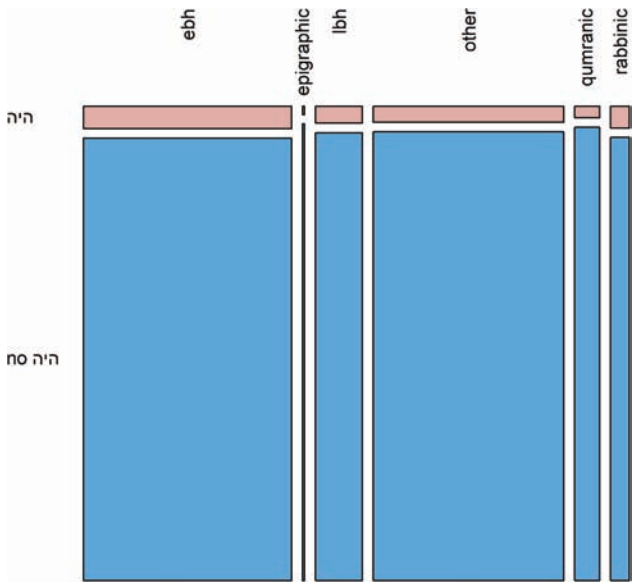

FIGURE 4.1 *Association of language phase and clause type for clauses with and without* היה

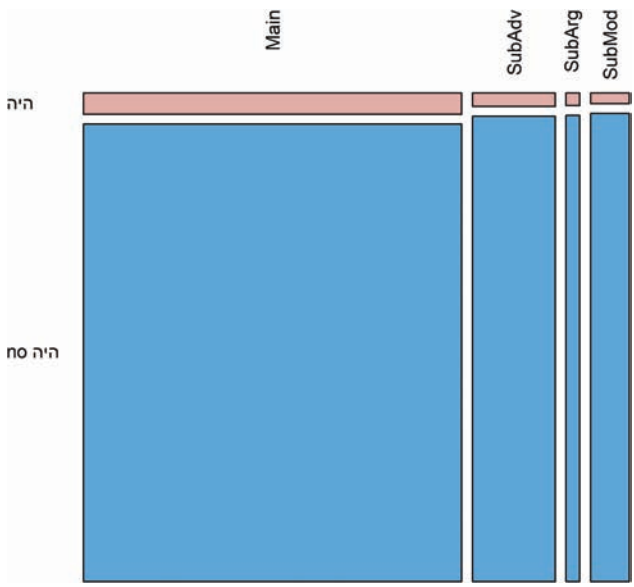

FIGURE 4.2 *Association of main and subordinate clauses and clause type for clauses with and without* היה

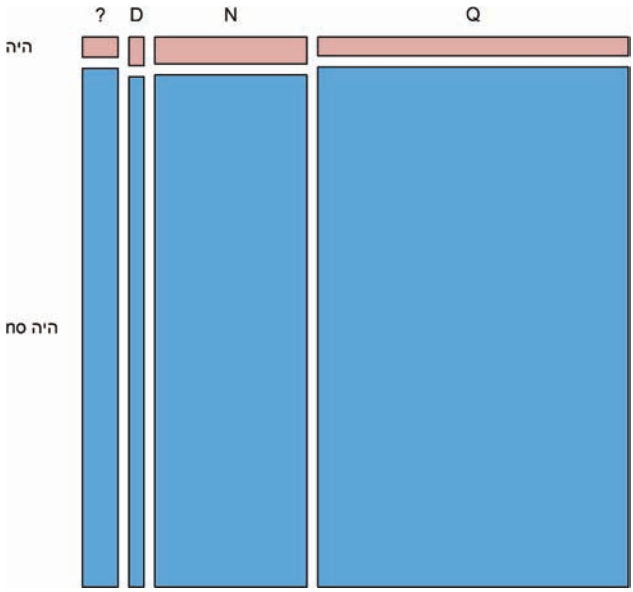

FIGURE 4.3 *Association of discourse type and clause type for clauses with and without* היה

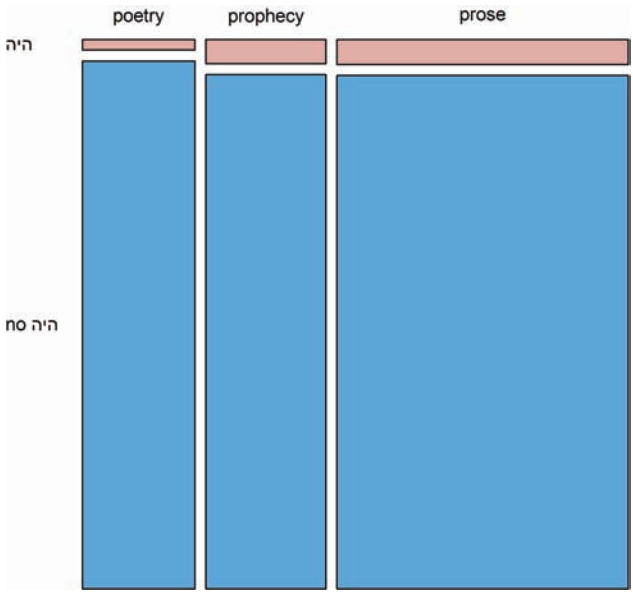

FIGURE 4.4 *Association of genre and clause type for clauses with and without* היה

proportions of the use of היה in different supposed language phases (in the case of figure 4.1) visually.

Obviously, the clauses containing היה are only a small subset of all the clauses in the corpus. The variation that can be found between different discourse types, genres, and main and subordinate clauses seems to cohere well with what is found in scholarly literature on the function of היה. In general, it is proposed that the verb adds TAM to the clause, and one would expect this in N-clauses in prose, because time generally plays a more important role in narrative than, for instance, in poetry. Also, היה occurs more in main clauses than in subordinate clauses, which is also as expected, because in general main clauses constitute the backbone of a story. In these clauses, tense, aspect, and mood are defined and inherited by subordinate clauses.

Of the different language phases, EBH has the highest concentration of clauses containing היה (4.8%), closely followed by Rabbinic Hebrew (RH) (4.6%). The concentrations in LBH (3.6%), QH (2.5%), and epigraphic Hebrew (1.7%) are lower. The relatively high concentration of היה clauses in RH is expected, because of the often recurring typically Rabbinic periphrastic construction of היה with a participle (הוא היה אומר, "he used to say", occurs often in Pirqe Avot, see Dyk (1984)). Based on the conventional diachronic model, one would expect a similarly high level of היה in LBH and QH if there were an increase of the periphrastic construction and other constructions would remain constant. This, however, does not seem to be the case here. It is beyond the scope of this research to investigate this issue in depth. Now we move on to clauses containing subject and predicate complements containing the verb היה.

Clauses with subject and predicate complement with and without היה

Clauses containing both a subject and a predicate complement often resemble each other. In both clause types, the subject can be a NP,[9] a PrNP,[10] a PPrP,[11] or a DPrP[12] and the predicate complement in both clause types can be a NP,[13] an AdjP,[14] or a PP.[15]

---

9 With היה: Gen 2:7, ויהי האדם לנפש חיה, "And the man became a living being"; Esth 2:5, איש יהודי היה בשושן הבירה, "There was a Jewish man at the citadel in Susa". Without היה: Gen 11:4, וראשו בשמים, "And his top (will reach) into heaven"; Isa 35:7, בנוה תנים רבצה, "In the haunt of jackals is its resting place".

10 With היה: 1 Sam 18:12, כי־היה יהוה עמו, "For the Lord was with him"; 2 Sam 8:2, ותהי מואב לדוד לעבדים, "And the Moabites became servants to David". Without היה: 1 Chr 4:42, ופלטיה ונעריה ורפיה ועזיאל בני ישעי בראשם, "And Pelatiah, Neariah, Rephaiah and Uzziel, the sons of Ishi are their leaders".

This section explores the association of various factors with the choice between using or not using היה in clauses with a subject and predicate complement. The main variables that will be investigated are the core variables of the Syntactic Variation project: discourse type, the genre of the text, main and subordinate clauses, and supposed language phase.

In the analysis, clauses containing יש and אין were removed, because these are considered to be different ways of expressing "to be". Consequently, other clauses containing a negation were also removed. Further, tripartite verbless clauses were removed.

The dataset consists of 8,883 clauses,[16] of which 7,706 are verbless clauses and 1,177[17] contain היה, which means that slightly more than 14% of all the clauses under consideration are היה clauses. These unequal frequencies of the two classes indicate that it is likely that in all or nearly all environments there will be a majority of verbless clauses.

Table 4.2 (see next page) shows the distribution of clauses with subject and predicate complement over the three sub-corpora of epigraphic Hebrew, BH, and QH. The total amounts of relevant clauses in each of the subcorpora differ greatly, the biblical subcorpus is the largest by far, but the percentage of היה clauses in the subcorpora is relatively similar: 13% (BH), 10% (QH), and 9% (RH).

---

11 With היה: Exod 4:16, ואתה תהיה־לו לאלהים, "And you will be to him as God"; Ezek 11:11, היא לסיר לא־תהיה לכם, "She will not be a pot for you". Without היה: Num 11:21, אשר אנכי בקרבו, "(The people), among whom I am"; Job 13:16, גם־הוא־לי לישוע, "This also will be my salvation".

12 With היה: Num 18:9, זה־יהיה לך מקדש הקדשים מן־האש, "This shall be yours from the most holy from the fire"; Ps 118:23, מאת יהוה היתה זאת, "This is the Lord's doing". Without היה: Qoh 6:2, זה הבל, "This is vanity"; 2 Chr 2:3, לעולם זאת על־ישראל, "This is forever on Israel".

13 With היה: Gen 1:2, והארץ היתה תהו ובהו, "The earth was formless and void"; Obad 1:18, והיה בית־יעקב אש, "Then the house of Jacob will be a fire". Without היה: Lev 13:22, נגע הוא, "It is an infection"; Ezra 8:13, ואלה שמותם, "These are their names".

14 With היה: Gen 36:7, כי־היה רכושם רב, "For their property had become too great"; Ruth 2:12, ותהי שלמה מעם יהוה אלהי ישראל משכרתך, "And your wages be full from the Lord, the God of Israel". Without היה: Song 1:5, שחורה אני, "Black am I"; Neh 8:11, כי היום קדש, "For the day is holy".

15 Examples in footnotes 9–12.

16 Or two datasets, one csv file contains the biblical data, the other contains the extrabiblical data. In R, they are merged together into one data frame. These are the numbers after some data cleaning.

17 This is about one-third of all the היה-clauses in the MT.

TABLE 4.2 *Distribution of* היה *clauses and verbless clauses in three Hebrew subcorpora*

	Epigraphic Hebrew	Biblical Hebrew	Qumran Hebrew	Rabbinic Hebrew
היה clauses	3	1113	34	27
Bipartite verbless clauses	18	7102	319	277

The following figures shows mosaic plots for the relationship between the clause types and the main independent variables under consideration, namely, the language phase, main and subordinate clauses, genre, and discourse type.

As expected, overall there are more verbless clauses throughout the studied phases, but the plot shows that היה-clauses occur more frequently in EBH than in LBH.

Figure 4.6 (see next page) shows the mosaic plot in which the association between clause type and main and subordinate clauses is visualized.

The first thing that can be noticed is the unequal proportion of main and subordinate clauses. The figure shows an underrepresentation of היה in subordinate clauses.

Figure 4.7 shows the association between clause type and discourse type.

This figure shows that the majority of relevant clauses can be found in quoted speech, Q. This is not surprising, because poetry and prophecy consist mainly of Q, and also in narrative texts there is a substantial number of Q-clauses. N and D have a relatively strong overrepresentation of היה-clauses.

Finally, the association between genre and clause type is shown in figure 4.8.

The figure shows that there is an underrepresentation of היה clauses in poetry.

In summary, there is an overrepresentation of היה-clauses in main clauses, discourse type N, prose, and prophecy. N can be found mainly in prose, so this joint overrepresentation might be no surprise. Furthermore, independent clauses have a stronger association with היה-clauses than do subordinate clauses. This is not surprising: most Hebrew grammars agree that it is used to provide a clause with TAM. If that is true, and if independent clauses are more related to the main line of a story, it is logical that they need היה as a carrier of TAM. It is interesting that היה is more strongly associated with EBH than with LBH, because there is no clear reason why this kind of clause would be more frequent in EBH than in LBH. Overall, these results are similar to the results found in figures 4.1–4.4, in which all clauses with היה were contrasted with all clauses without היה.

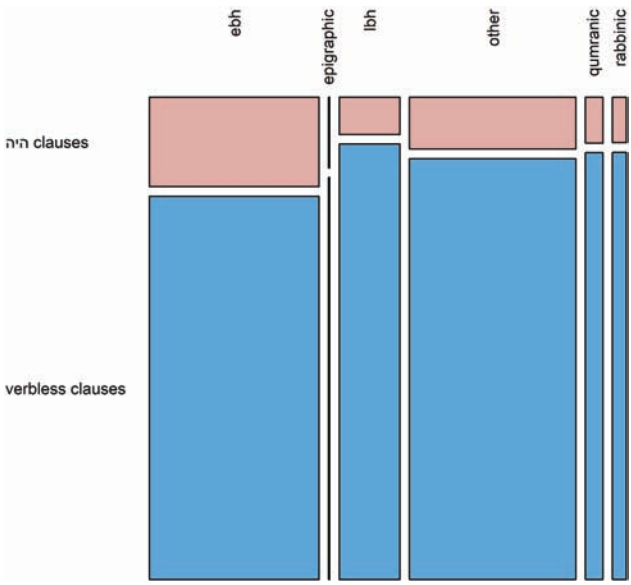

FIGURE 4.5 *Association of language phase and clause type for clauses with and without* היה

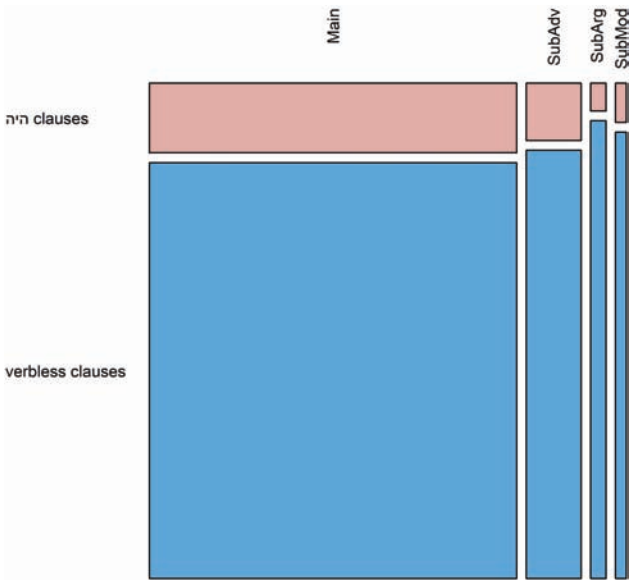

FIGURE 4.6 *Association of main and subordinate clauses and clause type for clauses with and without* היה

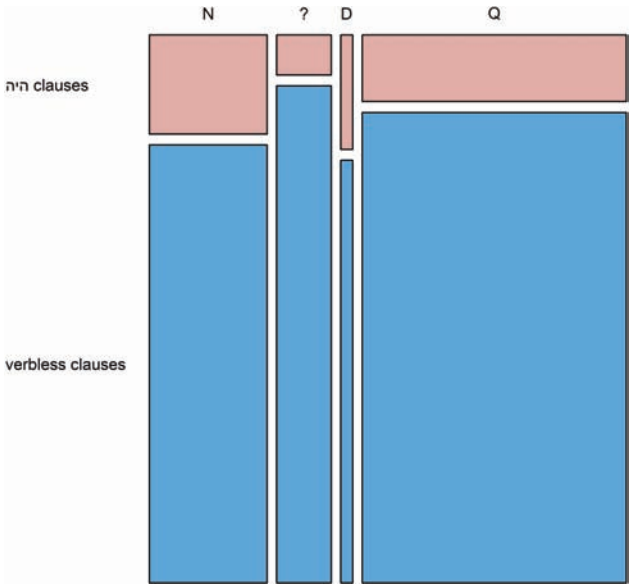

FIGURE 4.7 *Association of discourse type and clause type for clauses with and without* היה

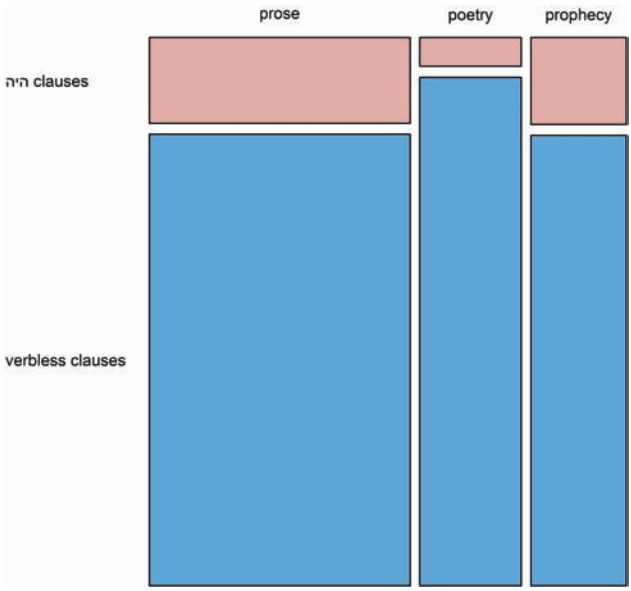

FIGURE 4.8 *Association of genre and clause type for clauses with and without* היה

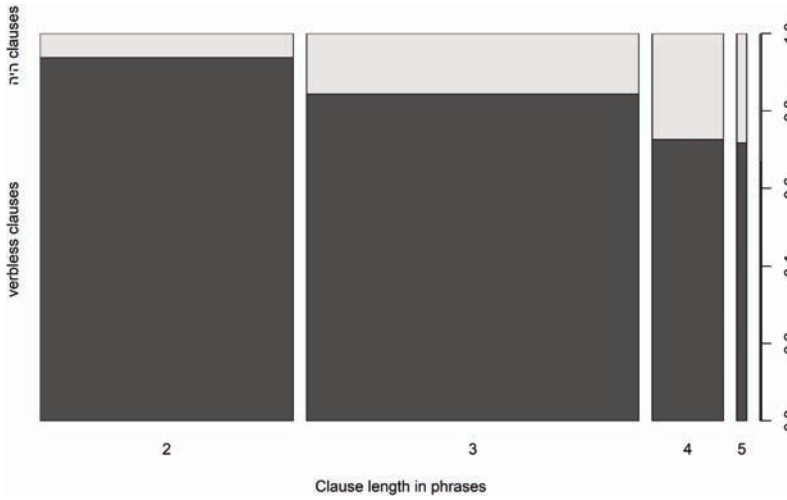

FIGURE 4.9 *Association of clause length and clause type for clauses with and without* היה

Clause length

Is the length of a clause related to the choice of using היה-clauses or verbless clauses? Figure 4.9 is a spine plot in which the relationship between clause length and clause type is shown.

Notice that the following clauses both have length 2:

דוד מלך
דוד היה מלך
David is/was king.

In היה clauses, the number of phrases is counted, except the verbal predicate. Figure 4.9 shows that there is a general tendency towards a more frequent use of היה in longer clauses. This tendency raises the question whether there are specific phrase types that associate with היה. In the dataset, there are columns for all the phrase types that occur in the clauses under consideration. 1 and 0 indicate presence or absence of this specific phrase function in the clause. Some of these phrase functions are rare, but the more frequent phrase functions are explored in the R-file.[18] On the basis of frequency

18 This is the file plots_and_analysis.R, in the folder https://github.com/MartijnNaaijer/phdthesis/tree/master/Ch4_Expressions_of_to_be/hyh_verbless.

and relevance, it was decided to include conjunctions, negators, question phrases, interjections, and time phrases in the model.

4.2.7. *Results*

Modeling variation

The exploration has shown a number of tendencies in the data, but if we want to know whether certain patterns are based on coincidence or not, a statistical test has to be performed with all the variables together. For this purpose, regression analysis is a good choice, because the results show the strength and significance of each of the independent variables in relation to the dependent variable.

Building the model

The modelling process has been done using forward variable selection. With this approach one starts with a model having only the random effect, and with each added variable one checks whether the model fits the data better than the simpler model.

In a multivariate regression model, the effects of the individual predictors are independent of each other. It is also possible to include interaction effects. In an interaction effect, the joint effect of two predictors is measured. In the case of a significant interaction effect, the joint effect of two variables differs from the sum of the individual effects, as is often the case in practice. In itself, it would be interesting to include interaction terms, but interactions with categorical variables produce large amounts of interaction terms which may be difficult to explain linguistically. Some exploration of interaction terms has been done, but for reasons of explainability and simplicity, I have decided to include main effects only.

The R-code that was used for the complete model is as follows:

```
bam.fit <- bam(cl_type ~ main_sub + genre + ebh_lbh + Conj + Intj +
Time + Ques + pc_type + cl_len2 + mother + s(cl_id2, book2, bs = 'fs', m =
1), data = dat.t, family = 'binomial')
```

Here bam.fit is the name of the model, and bam is the function from the mgcv package used here. cl_type is the dependent variable (left of the ~), the part s(cl_id, book2, bs = "fs", m = 1) is the factor smoother (bs = "fs"), and m = 1 limits the curviness per individual book. The link function is by default logit and the part family =

'binomial' is based on the structure of the dependent variable, which has two values. The independent variables are separated by plus signs.

In a regression model with categorical variables, one level of each variable is chosen as base level, and the p-value of each of the other levels indicates whether its value differs significantly from the base level. The variables in the model have the following values as base level:

main_sub: Main
txt_type: N
genre: prose
ebh_lbh: EBH
mother: nominal
pc_type: PP

The following variables indicate the presence of a phrase with that function in the clause. The basic levels are 0, which means that a phrase with that function is absent in the clause.

Conj: 0
Intj: 0
Time: 0
Ques: 0

Clause length is a numeric variable, so it does not have a basic level. Here it is called cl_len2, instead of cl_len. In this new variable cl_len2, the mean value of the clause length is set to zero, which makes values of the other variables easier to interpret.

Table 4.3 shows R's summary of the model. The first column shows the name of the level of a certain factor variable. For instance, genrepoetry gives the results of the value poetry of the variable genre, in relation to the basic value of that variable, which is prose. The intercept in the first row gives the result if all the variables are in their basic level. The following columns show, respectively, the effect size (called estimate here), the standard error, the z-value, the p-value, and, finally, the stars indicate at which level of alpha the effect is significant. In this research, an alpha value of 0.05 is used, which means that all the p-values lower than 0.05 are significant. These are marked with *, **, or *** in the last column. More stars mean a lower p-value and stronger significance.

TABLE 4.3 *Summary of the* GAMM *model*

	Estimate	Std. Error	Z-value	P-value	
Intercept	2.242778	0.216954	10.338	<2e-16	***
main_subSubAdv	0.964072	0.123119	7.830	4.86e-15	***
main_subSubArg	1.682743	0.265591	6.336	2.36e-10	***
main_subSubMod	1.181529	0.286560	4.123	3.74e-05	***
genrepoetry	1.178989	0.316733	3.722	0.000197	***
genreprophecy	0.180098	0.375089	0.480	0.631123	
ebh_lbhlbh	1.252875	0.460250	2.722	0.006486	**
ebh_lbhother	0.458538	0.378635	1.211	0.225884	
ebh_lbhepigraphic	0.271218	0.919575	0.295	0.768041	
ebh_lbhqumranic	0.833427	0.662275	1.258	0.208236	
ebh_lbhrabbinic	-0.562957	0.782589	-0.719	0.471924	
pc_typeAdjP	1.306309	0.141003	9.264	<2e-16	***
pc_typeAdvP	0.214582	0.247536	0.867	0.386013	
pc_typeDPrP	0.009393	0.930700	0.010	0.991947	
pc_typeInrP	13.137498	100.440766	0.131	0.895935	
pc_typeIPrP	2.425173	0.465058	5.215	1.84e-07	***
pc_typeNP	1.232721	0.083792	14.712	<2e-16	***
pc_typePPrP	13.346076	187.515521	0.071	0.943260	
cl_len2	-0.611704	0.070972	-8.619	<2e-16	***
motherimpf	-1.916307	0.136099	-14.080	<2e-16	***
motherimpv	-1.718744	0.213891	-8.036	9.31e-16	***
motherinfa	10.397227	388.880426	0.027	0.978670	
motherinfc	-1.618494	0.261082	-6.199	5.68e-10	***
motherno_mother	-2.273138	0.252665	-8.997	<2e-16	***
motherno_pred	-1.503936	0.172454	-8.721	<2e-16	***
motherperf	-1.815816	0.124878	-14.541	<2e-16	***
motherptca	-1.022856	0.209310	-4.887	1.02e-06	***
motherptcp	-0.524659	0.463484	-1.132	0.257639	
motherwayq	-1.595034	0.131541	-12.126	<2e-16	***
Conj1	-0.413500	0.107543	-3.845	0.000121	***
Intj1	1.651225	0.304443	5.424	5.84e-08	***
Time1	-0.617226	0.185661	-3.324	0.000886	***
Ques1	1.151467	0.361186	3.188	0.001433	**

Interpretation and visualization of the model

Levels that do not differ significantly from the base level of the variables were not merged with the base level, as is done with categorical variables sometimes. In this case, the intercept level is 2.242778. Since the model is a logit model, the odds are $e^{2.242778} \sim 9.4$. This means, that if all the predictors are in their basic level, verbless clauses occur 9.4 times as often as היה clauses.

In the case of a negative value in the column "Estimate" in table 4.3, there is an increase of the use of היה, relative to the base level of that variable. To calculate the logit of clauses with the value yiqtol (impf) in the variable mother, one has to add its value to the intercept level: 2.242778 - 1.916307. In this case, the odds are $e^{2.242778 - 1.916307} = 1.38$. This means that there is a strong increase of the use of היה in clauses of which the mother has a yiqtol verb, relative to the base level, in which the mother is a verbless clause. Likewise, the logit of an argument clause (level SubArg in variable main_sub) with a mother yiqtol can be calculated by 2.242778 - 1.916307 + 1.682743. The rest of the variables remain in their basic levels.

One of the main variables not present in the model is the variable discourse, of which the levels did not differ significantly from each other. The other variables in the model all have one or more levels that deviate significantly from the intercept level.

Model visualizations

In the following figures the model is visualized. First the presence or absence of phrases with a specific function are discussed, then the core variables of the Syntactic Variation project follow. In the figures, only the base levels and the levels with a significant effect are shown.

Figure 4.10 (see next page) shows the variation in the use of היה in the presence/absence of question phrases and interjection phrases. On the vertical axis the presence or absence in a clause of a phrase with this phrase function is shown by 1 (a phrase with this function is present) and 0 (a phrase with this function is absent). The horizontal axis shows the log-odds[19] of the level under consideration, with all the other variables on their base level, and the numeric variable clause length has its mean value.

19 Log-odds is the same as logit, see also Appendix B.

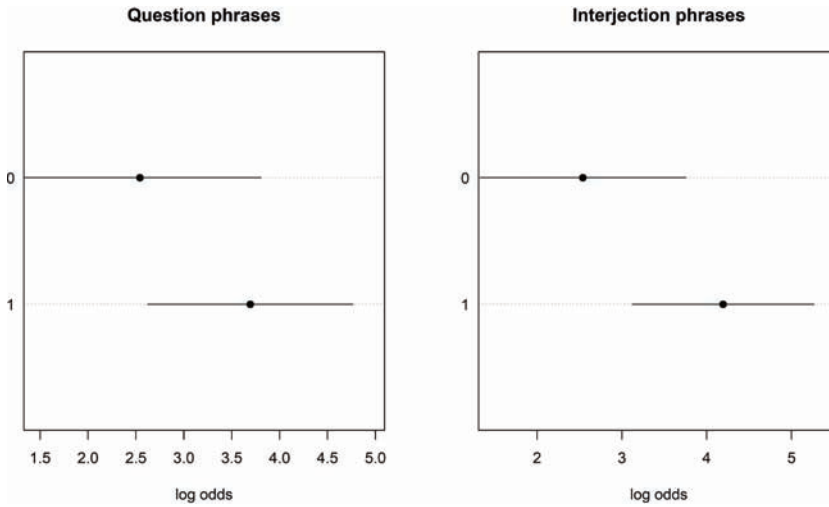

FIGURE 4.10 *The effect of the presence of question and interjection phrases on presence/absence of* היה

The horizontal line around the mean value shows the confidence interval.

The presence of question and interjection phrases both result in a strongly significant lower use of היה in these clauses, than in clauses without these phrases. Question and interjection phrases are both associated with quoted speech, and without these variables in the model the levels Q and N of the discourse variable differ significantly. This means that the difference between quoted speech and narrative with respect to the presence and absence of היה can be explained at least partly by the presence and absence of these two phrase types. This means also, that there seems to be no intrinsic difference between Q and N, apart from the concentration of question and interjection phrases.

The tendency is opposite in the case of the presence of conjunction phrases and time phrases. The presence of these phrases is associated with an increased use of היה. See figure 4.11 on the next page. The association of time phrases and היה does not come as a surprise. If the verb היה adds tense to the clause, then a time phrase can be added to specify that tense further. The associaton of conjunctions and the verb היה may be associated with a more complex, narrative style.

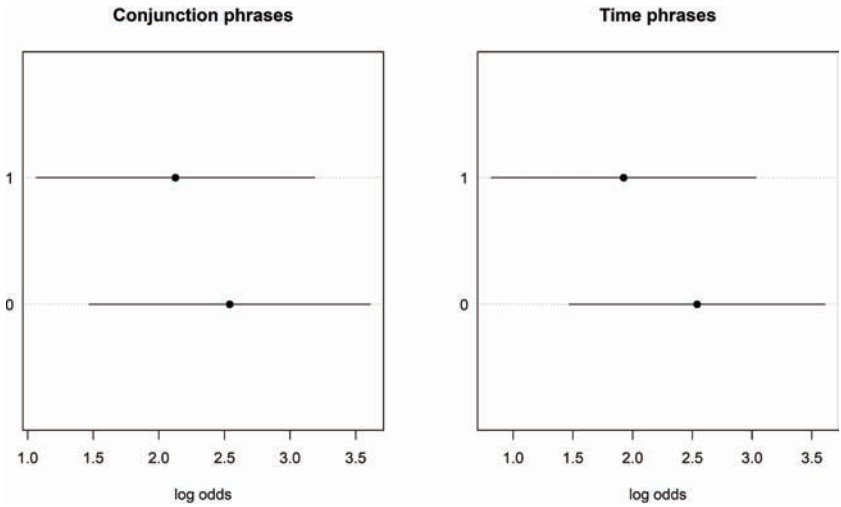

FIGURE 4.11 *The effect of the presence of conjunction and time phrases on presence/absence of* היה

The core variables of the Syntactic Variation project

The discourse environment of a clause did not end up in the model, because its main alternative value Q does not differ significantly from the basic level N. The exploration suggested some difference, but it turned out to be insignificant, as already explained. The variable language phase is included in the model. The use of היה in LBH differs significantly from that in EBH. It is the only level of this variable which differs significantly from the base level EBH.

Figure 4.12. shows the variation in the attestation of היה between the language phases.

The attestation of היה in the Rabbinic texts is similar to that in EBH, whereas especially QH and LBH have a lower use of היה. Only in LBH does the use of היה deviate significantly from the use in EBH. This does not mean that there is no significant variation at all between EBH on the one hand and QH and RH on the other hand, but the given data, with only a few Qumran and Rabbinic texts give no reason to state that they do differ significantly.

Figure 4.13 shows the proportions of verbless clauses and היה clauses in the EBH and LBH books separately.

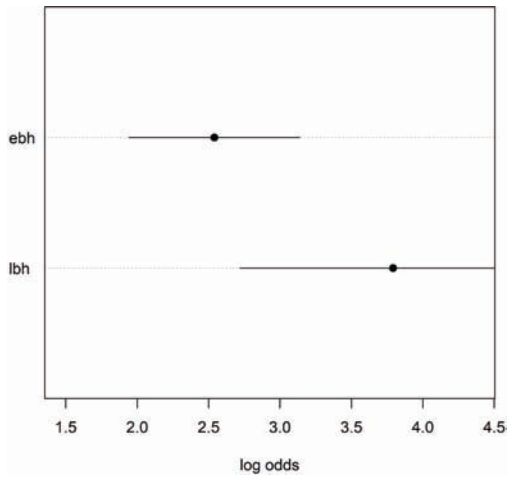

FIGURE 4.12 *The effect of language phase on the presence/absence of* היה

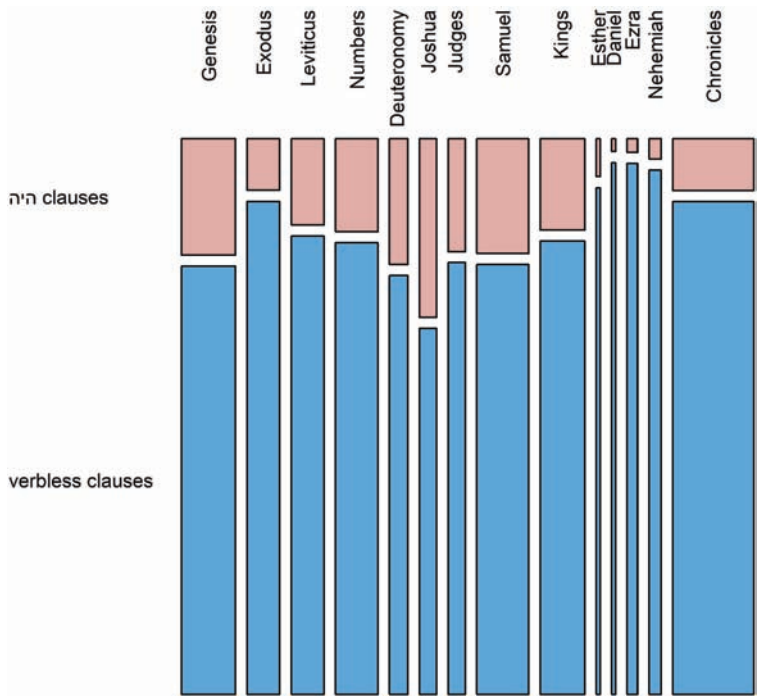

FIGURE 4.13 *Association of clause type and book in* EBH *and* LBH *books for clauses with and without* היה

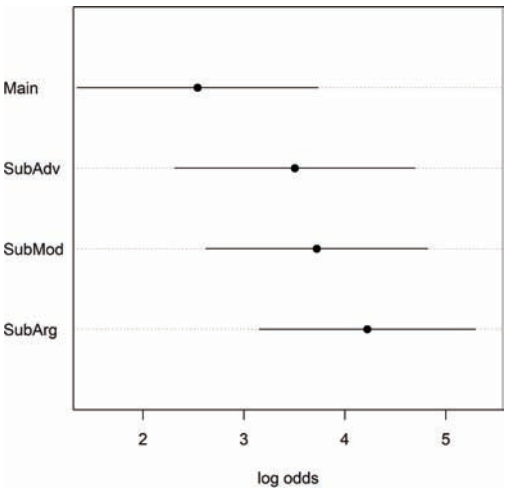

FIGURE 4.14 *The effect of main and subordinate clauses on presence/absence of* היה.

היה clauses are especially rare in the books of Esther, Daniel, Ezra, and Nehemiah, in which they occur, respectively 3, 2, 3, and 5 times. The book of Chronicles has a higher density of היה clauses, but still, the figure shows that this density is lower than in the EBH books. Discussions of the use of היה in a diachronic perspective focus on the periphrastic construction (היה + participle), which has an increased use in LBH and RH. It is tempting to see this, in relation with the decrease of היה in constructions with a subject and nominal predicate complement, as discussed in this research, as a shift in the use of היה. It could indeed be the case that such a shift is visible in Hebrew, but a remark needs to be made. The data contains only a relatively small amount of post-biblical material, so later research, based on a substantial amount of data, may give entirely different results.

Figure 4.14 shows the variation in the choice of היה or a verbless clause between main clauses and the three types of subordinate clauses. In the variable main and subordinate clauses, the main clauses have the highest use of the verb היה, and the subordinate clauses all deviate significantly form the main clauses, all with a strongly reduced use of היה. This confirms the impression that the exploration had given already.

Clauses on the argument position of another clause (SubArg) and clauses functioning as a modifier within a NP (SubMod) have the lowest probability of having היה. This is not surprising, because the tense, aspect, and mood of a clause which is the subject or object of another clause retrieves this from the main clause.

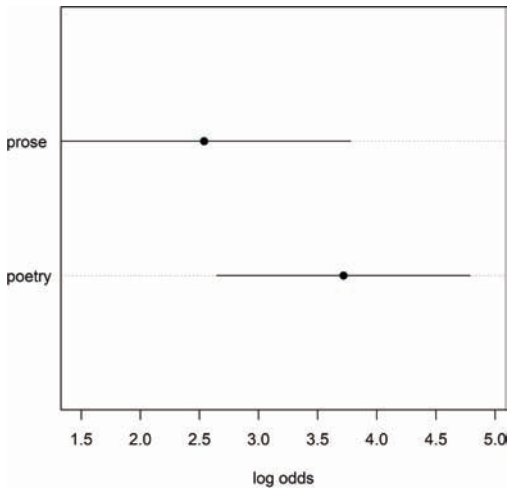

FIGURE 4.15 *Effect of genre on presence/absence of* היה

Genre is the last of the main variables of the project in which there is significant variation between its levels, as in figure 4.15.

Clauses in the basic level, which is prose, have a significantly higher probability of containing היה than clauses in prophecy and especially poetry. This may be related to the narrative character of prose. References to past events occur often in narrative prose, so it is not surprising that as the carrier of tense morphology, the verb היה, occurs mostly there, instead of in "timeless" poetry.

Figure 4.16 shows the effect of clause length on the use of היה.

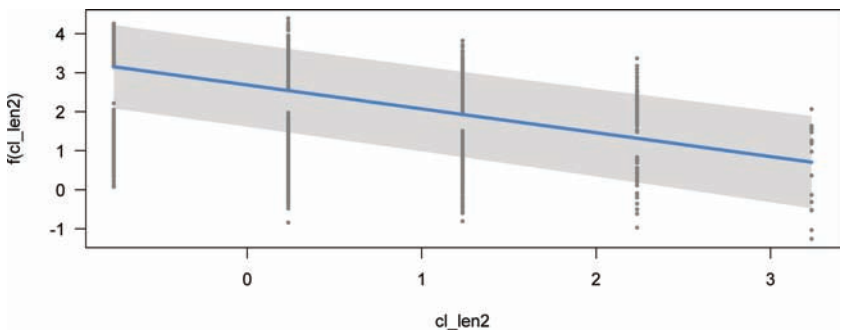

FIGURE 4.16 *Relationship between clause length and the use of* היה

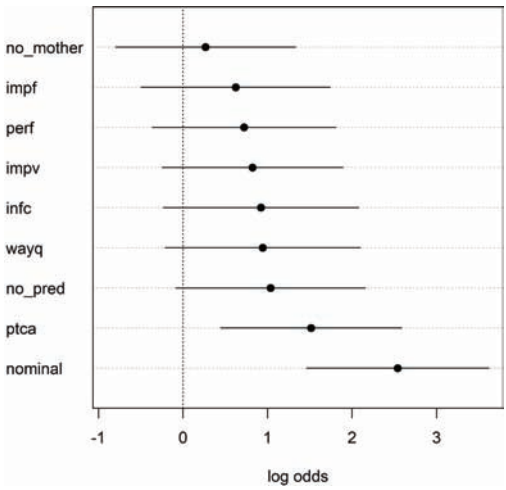

FIGURE 4.17 *Effect of the tense of the mother of a clause on the presence/absence of* היה

Longer clauses tend to have a higher probability of having היה. It had become clear already that the presence or absence of certain phrases is associated with היה-clauses or verbless clauses, but independent of these effects comes the effect of clause length. It is surprising that the effect of clause length has not disappeared with the presence of the various different phrase functions in the model (Conj, Ques, Time, and Intj), because the presence of these phrases causes the clause to be longer. There is no obvious semantic explanation for the increased use of היה in longer clauses, but its function may be related to what is called the "copula of separation" (Driver 1892a: 270 n. 4). In this interpretation, the copula does not have a meaning, but it is there to give structure to the clause.

Figure 4.17 shows the effect of the tense of the mother of a clause and the use of היה. In this figure, the value nominal is the base level of the variable. It has the highest log odds, which means that it has the lowest probability of clauses with היה. Nearly all the other levels have a significantly higher probability of using היה in the clause, and the effect size is high. The figure suggests that the levels of the variable mother can be divided in two groups. On the one hand there is the level nominal, in which case there is a high probability of finding a verbless clause, and on the other hand there are the other levels. In the case of the other levels, most values concern mother clauses with a finite verb. In these cases, the probability of finding היה in a clause is still lower than 0.5, but substantially higher than in the case of a verbless clause as mother. This gives the impression that verbless clauses occur

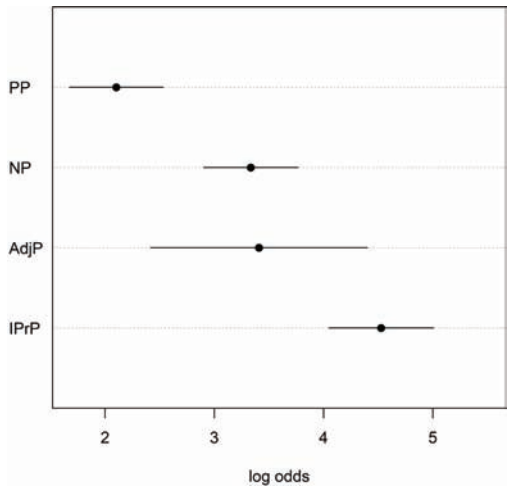

FIGURE 4.18 *The effect of the phrase type of the predicate complement on the presence /absence of* היה

mainly when the mother of the clause is also a verbless clause, so there is a sequence of verbless clauses. היה clauses occur more often if the mother clause also contains a verb, or if the clause has no mother. This is the case if a new hierarchy begins.

The base level of this variable phrase type of the complement (figure 4.18) is PP, prepositional phrase. The figure shows the levels of PP and the levels deviating significantly from PP. These are AdjP (adjective phrase), NP (noun phrase), and IPrP (interrogative pronoun phrase). Clauses with such a predicate complement type have a significantly lower use of היה in constructions with a subject and predicate complement. This difference in the behavior of clauses with a PP and other predicate complements may well be explained by semantic variation of the verb between these clauses. As was suggested by various authors, היה is not only used as a simple copula, but can have a meaning like "to happen", or "to become", and this meaning is likely to occur more in clauses with a PP as predicate complement.

This does not mean that this semantic difference always occurs between clauses with and without PP as predicate complement, but there is a clear tendency. This is one of the advantages of using statistical techniques in analyzing language. In grammars, generally only examples of a certain linguistic phenomenon are given. These examples may make clear that there is structural or semantic variation, but not which global tendencies there are.

Figure 4.19 shows the smoother of the model.

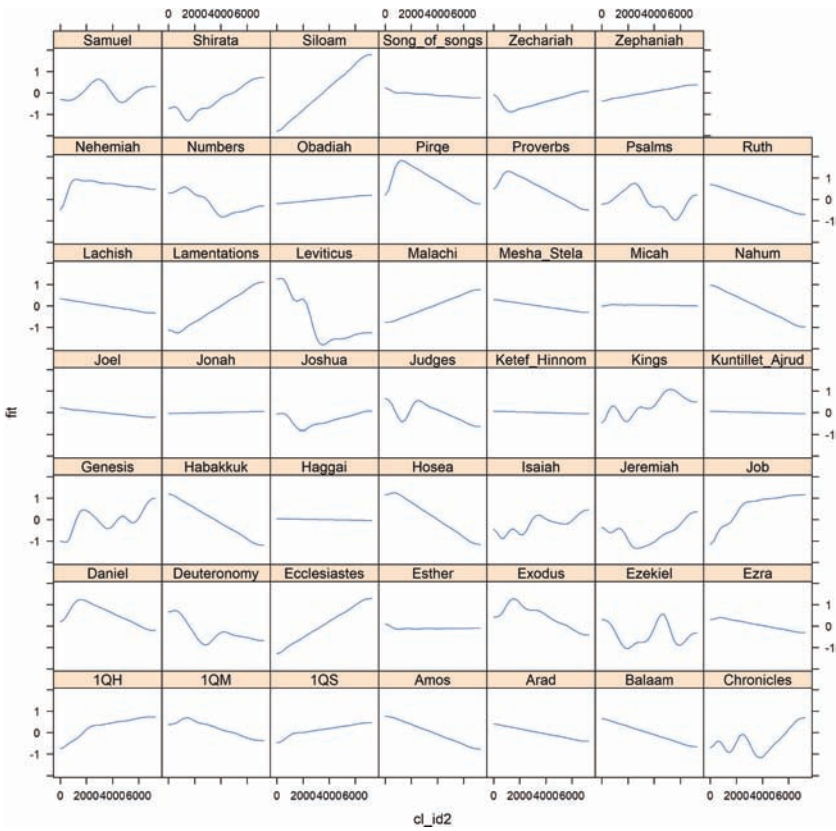

FIGURE 4.19 *Model smoother*

These smoothers show the trend throughout each book in the dataset, and function as random effect in the model. The y-axis shows the log-odds[20] of the probability of finding a verbless clause, on the x-axis the place in a book is represented, using the Text-Fabric word identifier. As can be seen, there can be quite a bit of fluctuation within a book.

The explained deviance of the model is around 27%. This is modest, but on the other hand, by basing the analysis on mostly formal criteria it is possible to uncover a number of variables that are associated with the clause type. The Hebrew Bible is an ancient corpus. Little is known about its background from direct evidence. I

20 In this figure, it is centered around zero.

have modeled the variation in the use of היה clauses and verbless clauses on the basis of mostly formal characteristics in the structure of clauses and texts and the model shows a number of interesting tendencies.

Discussion of the results

On the basis of the literature one would expect that an important part of the variation between verbless clauses and היה clauses can be explained by the insight that היה adds TAM to the clause. This is confirmed by the results of this research. The clearest sign of this is that there is a significantly higher use of היה in clauses containing a time phrase.

The main levels of the discourse variable, Q and N, do not differ significantly from each other in the use of היה, but there are various signs that they do differ in an indirect way. Clauses containing interjection phrases and question phrases, have a significantly reduced use of היה, while there is an increase of the use of היה in clauses containing conjunctions. Interjections and questions are more related to the domain of Q, while conjunctions, leading to more complex sentence structures, are more related to N.

In LBH, there is a significant decrease of the use of היה, as compared to its use in EBH. Can this be seen as a sign of diachronic variation? The traditional diachronic approach assumes a direction of change in which there is an increased use of late language going from EBH to LBH, and from there it increases further to QH and RH. This is clearly not the case in the variation between היה clauses and verbless clauses. LBH and QH have a lower use of היה than EBH, and of these only LBH's use differs significantly from EBH. RH is more similar to EBH than LBH is. Of course, the amount of data used for QH and RH is limited, so these observations have limited value. Despite this, the use of היה in clauses with a subject and predicate complement is consistently lower than in EBH texts. It is possible that during some time period, at a certain place there was a reduced use of היה in this kind of constructions. This could be investigated in a broader study on the use of היה in Classical Hebrew. There are other signs of a shift in the use of היה in LBH. Various scholars point to an increased use of the periphrastic construction in LBH (e.g., Van Peursen 2004: 226, others are skeptical, e.g. Muraoka 1999: 195). An increased use of the periphrastic construction could be related to a decreased use of היה in the constructions studied in this section. On the other hand, the increased use of the periphrastic use in LBH is usually linked to its frequent use in QH and RH. On the basis of the data used in this study, the use of היה in constructions with subject and predicate complement in QH and RH is more similar to EBH.

Although there is a significant difference between EBH and LBH, figure 4.13 shows that there is substantial variation in the use of היה between individual books. This may be caused by a variety of factors. Some of these may be among the variables discussed in this section, others may be related to variation in subgenres. The genres discussed in this research, prose, poetry, and prophecy are broad, and could be split in various subgenres.

Interjections and question phrases are typical of clauses in direct discourse. However, in the present model, the feature txt_type, which represents the discourse embedding of a clause, is removed because of the lack of significant variation between the different levels. In a GAMM-model, in which txt_type is the only fixed effect, the levels Q and N differ significantly, but in the presence of the features Conj, Rela, Time, Ques, and Intj the effect of the discourse type disappears. Apparently, the difference between N and Q in relation to the difference between verbless clauses and היה clauses can be found at least partly in clauses with these phrases.

Longer clauses have a significant tendency to use היה more often. The most obvious reason for this is that היה gives structure to the clause.

Looking at the phrase type of the predicate complement, it is striking that if the predicate complement is a PP, היה is used significantly more than with other phrase types. It is possible that this is related to the meaning of the verb. In general, היה is translated with "to be", but in these cases a possible translation with verbs like "to happen" or "to occur" might be possible.

The results suggest a number of ideas for further research. As discussed in the review in chapter 2, Polak argues that there are various styles of Hebrew with varying sentence complexities. In this research, interactions between variables have not been studied, but it would be interesting to study the relationship of the combination of clause length and language phase on the use of היה. It is surprising that there is no significant variation between the levels of the variable discourse type.

In summary, based on mainly formal characteristics of a clause, we can conclude that the use of היה is conditioned by a variety of factors. Most of these seem to be related to the idea that היה adds TAM to the clause, but there are also other valid explanations.

4.3. Bipartite verbless clauses with and without the particle יש

4.3.1. *Problem and research question*

The Hebrew particle of existence, יש, can be found in a variety of verbless clauses. יש can occur without subject or predicate complement, e.g., in 1 Sam 9:12 יש, "It is". It can also occur with a PP predicate complement, e.g., Prov 3:28, ויש אתך, "It is with you", or with a NP subject alone, e.g., Job 11:18, כי־יש תקוה, "Because there is hope". There are also cases of יש with an enclitic subject and a participle, e.g., Gen 43:4, אם־ישך משלח את־אחינו אתנו, "If you send our brother with us". In this section, the most frequently occurring clause construction with יש, having a subject and a PP predicate complement, is studied. In these constructions, the particle indicates the existence of something or someone in a specific place (Muraoka 2013).

Existential clauses containing the particle יש are semantically similar to their counterparts without the particle, as Muraoka indicated. Muraoka doubts that there is any functional difference between the following clause and its equivalent without יש:

Gen 18:24 אולי יש חמשים צדיקם בתוך העיר

In both cases, a valid translation would be "Perhaps there are fifty righteous people within the city" (Muraoka 2013). Although Muraoka has described some characteristics of clauses containing יש, he and others have not made a systematic comparison of clauses with and without יש. Therefore, the research questions of this section are:

> Is it possible to distinguish between clauses with and without יש on the basis of formal characteristics of the relevant clauses? And if so, which are the most important predictors? What does this mean for the use of the particle?

The clauses that are considered contain an indefinite NP or IPrP subject and PP predicate complement. Clauses with this structure are the most frequent complete clauses with יש in the MT.[21] In the section on היה clauses and bipartite verbless clauses,

---

21 יש occurs 138 times in the MT. In 50 cases, the clause has an indefinite NP or IPrP as subject and a PP as predicate complement. There are only nine cases with a definite subject and PP predicate complement:

Gen 28:16, אכן יש יהוה במקום הזה, "Surely YHWH is in this place"; Gen 44:26, אם־יש אחינו הקטן אתנו, "If our little brother is with us"; Exod 17:7, היש יהוה בקרבנו, "Is YHWH among us?";

TABLE 4.4 *Frequencies of bipartite clauses with and without* יש*. All these clauses contain an indefinite subject and a* PP *predicate complement*

	Biblical texts	Extrabiblical texts
Clauses without יש	1076	177
Clauses with יש	50	17

a regression model was used to analyze the variation between these clause types. In the research on variation between clauses with and without יש, this would also have been a good option, but there are some differences between the datasets that prevent the regression model from being the best tool to analyze these data. In the first place, there are fewer relevant clauses for the study of יש clauses than for the study of היה clauses. See table 4.4.

In section 4.2, the dataset contained more than 1,100 clauses with היה, but in the case of יש, there are only 67 clauses available.

In the second place, there is a stronger imbalance between the classes of the output variable. The יש dataset contains only 67 clauses containing יש and 1253 clauses without יש. This means that there are nearly 19 clauses without יש for every clause containing the particle, while this proportion is about 1:7 (with versus without היה) in the היה dataset. The combination of unequal class sizes and a relatively small amount of data for the smallest class is a problem for a binomial regression model. Therefore, I have decided to use two approaches that are more robust for imbalanced class sizes. The algorithms of choice are Extreme Gradient Boosting (XGBoost), which is implemented in the R package xgboost,[22] and Random Forest, as implemented in the R package randomForest[23] (Liaw and Wiener 2002). Gradient Boosting and Random Forest are relatively new developments, which have not yet been applied in the study of BH.[24]

Judg 6:13, וישׁ יהוה עמנו, "And YHWH is with us"; 1 Sam 9:11, היש בזה הראה, "Is the seer here?"; 2 Kgs 3:12, יש אותו דבר־יהוה, "The word of YHWH is with him";

2 Kgs 10:23, פן־יש־פה עמכם מעבדי יהוה, "Lest there are servants of YHWH with you";

Jer 27:18, ואם־יש דבר־יהוה אתם, "And if the word of YHWH is with them"; Job 16:4, לו־יש נפשכם תחת נפשי, "If your soul were under my soul".

22 https://cran.r-project.org/web/packages/xgboost/xgboost.pdf.

23 https://cran.r-project.org/web/packages/randomForest/randomForest.pdf.

24 At least, as far as I am aware of. It was suggested by Forbes (2016: 922–923).

4.3.2. *Review*

Generally, clauses containing יש and אין are called existential clauses, because they express the existence or non-existence of something. Even though the particle אין is not investigated in this research, in scholarly literature the particles are often treated together.

In the relevant literature, existential clauses are not always described in the same way. For some, an existential clause is simply a clause containing יש or אין (for instance, Van Hecke 2005: 61). Geiger (2012: 367) mentions that some clauses with הנה are also existential clauses. Muraoka distinguishes between clauses describing absolute existence and existence in a certain location (Muraoka: 2013). These clauses may or may not contain the existential particles יש and אין. According to him, the locative meaning is the first meaning of the clauses containing these particles, and by extension it has the meaning of existence as such (JM, §154k). In a later publication, Muraoka asserts that יש clauses can only be existential, and not locative clauses (Muraoka 2013), even when the clause contains a prepositional phrase, as in the aforementioned example in Gen 18:24.[25] In this clause, it is not so clear that omission of יש would have changed the meaning of the clause, but in clauses in which there is no locative and the presence of something or someone is predicated, יש is obligatory, according to Muraoka. He gives the following examples:

> Prov 19:18 כי־יש תקוה
> For there is hope.

> 1Kgs 18:10 חי יהוה אלהיך אם־יש־גוי וממלכה אשר לא־שלח אדני שם לבקשך
> As the Lord your God lives, there is no nation or kingdom to which my lord has not sent to seek you.

> 1Sam 21:5 כי־אם־לחם קדש יש
> There is only holy bread.

25 As observed by Muraoka, clauses with this structure containing the particle יש indicate existence rather than location. The same is true for clauses without the particle. See, for instance the following examples: Gen 8:9, כי־מים על־פני כל־הארץ, "For water was on the face of the whole earth"; Gen 42:16, האמת אתכם, "Whether there is truth with you". In both these clauses, the issue is not where a thing (water, truth) is, but that it is at a certain place. This is generally the case if the subject is indefinite.

In Emphatic Words and Structures in Biblical Hebrew, Muraoka (1985: 77) states that יש and אין "are the exact Hebrew counterparts of the Indo-European copula", although he follows Joüon's opinion that both יש and אין have a certain volitive force, which Joüon explains in detail by referring to the cases in which יש is used in the protasis of conditional sentences in which a participle follows. However, according to Muraoka (1985: 77–78), the latter meaning stems from the context in which the clause occurs.[26] He remarks that the use of the positive particle יש is pleonastic, which is a formal indication of its emphatic meaning (Muraoka 1985: 79). The use of אין is not pleonastic, because it is an essential element of a negative verbless clause. For Muraoka, this is a sign that אין and יש are not simply symmetrical opposites, but that they must be treated separately. This is strengthened by the fact that there are cases of אין + participle that do not occur in a conditional clause, which is important to Joüon's argument. Also, there are cases of יש having the volitive force outside conditional clauses, and יש is only rarely accompanied with a definite NP, whereas אין occurs often with a definite NP. With the help of cognate languages (Akkadian, Aramaic), Muraoka shows that it is "almost incontestable" that the construction of יש + participle has a certain asseverative-confirmative nuance (Muraoka 1985: 77). This nuance contrasts with the views of Brockelmann, König, and Nöldeke, according to whom יש is nothing more than a mere copula. In later work, Muraoka stresses that יש is never used as a copula in an equational way in clauses like דוד מלך or מלך דוד, because in those cases היה is generally used.[27]

Within the group of existential clauses one can distinguish between clauses expressing existence and presence of a person or thing. In the case of existence יש, clauses express the existence of an impersonal or indefinite subject or the existence or presence of a person or thing (Van der Merwe, Naudé, and Kroeze 2004: 321). In the latter case, the subject is accompanied by a locative or possessive predicate (Baasten 2000: 1).

Structure of יש clauses

In two publications, Baasten describes the distribution of patterns occurring in clauses containing יש in the non-biblical DSS (Baasten 2000; 2006: 212–223). He focuses mainly on formal characteristics of the clause, following the work of Hoftijzer, who had a similar approach in his work on verbless clauses (Hoftijzer 1973). This means that

26 He gives the five examples in the MT: Gen 24:42, 49, 43:4, Deut 13:4, Judg 6:36.

27 And, likewise, not in the rarer patterns דוד הוא מלך and דוד מלך הוא, Muraoka (2013).

the function of a phrase (subject, object, predicate, et cetera) gets a less prominent role in his description, although it is not completely absent (for instance, Baasten 2000: 3). The focus is on determination and type of phrases. Baasten distinguishes the following four main patterns in יש clauses in the non-biblical Dead Sea Scrolls (Baasten 2000: 2–4. These patterns are read form left to right):

1. יש NPi

In this pattern, יש is followed by an indeterminate NP. It occurs twice in the non-biblical DSS, both times with question-ה prefixed to יש.

2. יש NPi PP
3. יש PP NPi

In these patterns, יש is followed by an indeterminate NP and a PP. These patterns are the most attested in the corpus under investigation. Pattern 2 occurs in stereotypical expressions introduced by כי (Baasten 2000: 2). In pattern 3, the PP precedes the NP. Here, the PP is a locative PP or consists of את with a possessive suffix. Baasten observes no clear functional difference between patterns 2 and 3.

4. יש PP

This pattern occurs once in the non-biblical DSS (11QT 58: 3–4).

In his PhD thesis, Baasten points out that subordinate יש clauses introduced by אשר always occur in prose, while those introduced by כי are all in poetry, but the reason for this is probably not linguistic (Baasten 2006: 212).[28] Most cases found in poetry are stereotypically recurring clauses like ואדעה כי "and I know, realise, understand, that …". Of the cases introduced by אשר, he observes that if אשר is not coreferential with the subject noun phrase, the order in the clause is NP PP אשר יש, of which he gives four examples (Baasten 2006: 212–213). In one case (4Q266 6 i 7–8), the PP follows the NP, but there the NP is definite.

However, the corpus investigated by Baasten contains the particle יש only 17 times, and this limits the value of the conclusions he described. The formal approach is valuable, because it makes it possible to compare similar patterns more easily.

28 On p. 214, Baasten gives five examples, all in the Hodayot: 1QH[a] 11:20–21, 14:6, 17:14 (2 times) and 22.

Van Hecke studied the order of constituents in existential clauses containing both a NP and a PP (Van Hecke 2005). In both BH and QH, he found that generally the NP precedes the PP if the PP consists of a preposition plus noun phrase and the PP precedes the NP if the PP consists of a preposition plus pronominal suffix.[29] Van Hecke did not find a clear semantic distinction between the different orders, but argues that it is possible that the shorter PP s consisting of preposition plus pronominal suffix tend to come earlier in the clause because of their length, which is a well-known cross-linguistic phenomenon (Van Hecke 2005: 69–70). There is a number of exceptions to the rule formulated by Van Hecke, and he explains these cases by a range of factors that influence constituent order in other clause types, such as the semantic properties of יש or the preposition governing the PP.

The particle יש and diachrony of BH

In "Syntax der Althebräischen Inschriften", Andreas Schüle (2000) observes that the particle יש does not occur in the pre-exilic Hebrew inscriptions (Schüle 2000: 219), in contrast to אין, which occurs 5 times in this corpus (Schüle 2000: 218). Although this absence of יש might be a coincidence, it could have diachronic significance (Schüle 2000: 219–222). He finds evidence for this in the Bible—יש is found mainly in later texts, like the Joseph story—and also in the development of particles and verbs of existence in other Semitic languages (Akkadian and Aramaic).

4.3.3. *Ensemble techniques: Random Forest and Gradient Boosting*

Predicting the clause type

As in the section on היה clauses and bipartite verbless clauses, the goal is to find out how the predictors are associated with the use of a word, in this case the particle יש. This is done here by predicting whether a clause contains the particle יש, using two machine learning techniques, Random Forest and XGBoost. These techniques come from a branch of machine learning called supervised learning, in which a model is trained that maps predictors to an output variable. If the output variable is a categorical variable, it is called the label of an observation. The function is learned by processing many examples of observations. With the learned function, also called a

29 This is a statistically significant effect with $p < 0.001$ (Van Hecke 2005: 65 n. 22).

model, one can predict the labels of new observations of which the label is unknown. In supervised learning, the word "prediction" does not necessarily have to do with an event in the future. In this research, predictions are made about the presence or absence of the particle יש on the basis of the main variables of the Syntactic Variation project and a number of other variables.

In supervised learning, it is crucial that the model is able to make correct predictions on unseen data. Often a model reaches high accuracy on the data on which it was trained, but this accuracy decreases when predictions are made on new data. In that case, the model does not generalize to unseen data, and the model is trained on idiosyncracies in the dataset and only to a lesser extent on the general patterns in the data. This situation is called overfitting and should be avoided. An important technique which helps to prevent overfitting is k-fold cross validation, which is used in this research. For an explanation of k-fold cross validation, see Appendix C.

One wants to obtain a prediction accuracy that is as high as possible, but in general it is not possible to reach 100% accuracy. Sometimes one misses important predictors, which is likely in the case of ancient data. The prediction itself is not the main goal of this study, but rather it is a means to find out which variables are important in the prediction and which variables are less important. This way one can get an impression of what conditions the use of the particle.

Decision trees and Random Forest

The first algorithm used for the classification of clauses is Random Forest, which is based on single decision trees. An example of a single decision tree can be found in figure 4.20.[30]

The figure shows how the species of an iris is predicted, based on certain properties of individual flowers. For instance, if we find a new iris flower and we wonder which species it is, the model on which figure 4.20 is based can be used, supposing that the species of the new flower is taken into account in the training set. If the petal length is ‹ 1.9, the model will predict that the species is "setosa". If the petal length is ‹ 1.9 and the petal width is ‹ 1.7 the model predicts that the species is "Virginica".

The advantage of such a tree is that its structure is intuitively understandable. In every node of the tree, a decision is made on the basis of some input variable, with

30 The tree is based on the "iris" dataset, a standard dataset in machine learning available in R. Without further preprocessing the tree was made with the function ctree in the R package "party" using Species as output variable and the other variables as predictors.

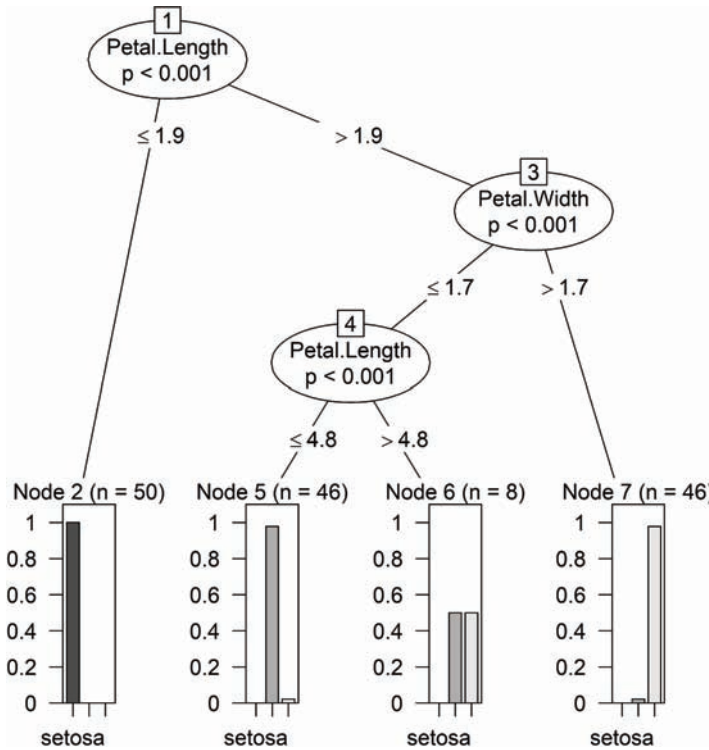

FIGURE 4.20 *Decision tree model based on the iris dataset*

the strongest predictors situated higher in the tree. In the leaf nodes, one can see the probabilities of each class on the basis of the training set. If one wants to make a prediction using a new sample, one can follow the scheme of the tree based on the predictor values in the sample, and the predicted class is the majority class in the leaf node. The algorithm builds the tree in such a way that the leaf nodes are as pure as possible, using entropy as a means to measure purity of subsets of the data.

These decision trees are prone to overfitting. Overfitting is reduced in Random Forest by making many independent trees, all being trained on a different random sample of predictors, which all overfit the data in a different way. The final classifier is averaged out by majority voting on the basis of predictions made by the separate trees. Algorithms that are based on classifying data using many models together are called "ensemble techniques".

Random Forest is a popular learning algorithm, and one of the reasons for this is

that one does not need much preparation of the data to apply it. It is implemented in various R packages. In this research, the package randomForest is used.[31]

Extreme Gradient Boosting

The second algorithm used in this section is Extreme Gradient Boosting (XGBoost). Boosting is a family of algorithms in which a number of weak learners together form a strong learner. A weak learner is a classifier that has a weak correlation with the true class. Kearns and Valiant (1989) asked the question whether more weak learners can together form a strong learner. This question led to the development of boosting, which nowadays is used in various applications. Gradient boosting is an iterative algorithm to which new models are trained to correct mistakes made by existing models, by giving wrongly classified samples a higher weight in the new model. It uses gradient descent for optimization and it is based on decision trees, just like, for instance, Random Forest. The iterative nature of the algorithm is an important difference: in Random Forest the trees are trained independently of each other, in XGBoost later trees depend on the ones that were made earlier.

XGBoost is implemented in the R package xgboost,[32] but it is also available for Python and other programming languages.[33] Extreme Gradient Boosting is fast, it outperforms many other algorithms on structured data and can be tuned with various hyperparameters, which makes it one of the favorite algorithms in data science competitions and research. For more information on Random Forest and XGBoost, see Appendix D.

4.3.4. *Data preparation and variables*

The dataset consists of the verbless clauses containing an indefinite subject and a prepositional phrase as predicate complement in the Hebrew Bible, 1QS, 1QH^{a}, 1QM, Pirqe Avot, and Shirata, as available in Text-Fabric.[34] This means that the there is overlap between the clauses without יש in this dataset and clauses without היה in the dataset used in section 4.2.

---

31 https://cran.r-project.org/web/packages/randomForest/randomForest.pdf.

32 https://cran.r-project.org/web/packages/xgboost/xgboost.pdf.

33 https://github.com/dmlc/xgboost.

34 The early inscriptions are present in the dataset, but they are not taken into account in the dataset, because יש does not occur in this corpus.

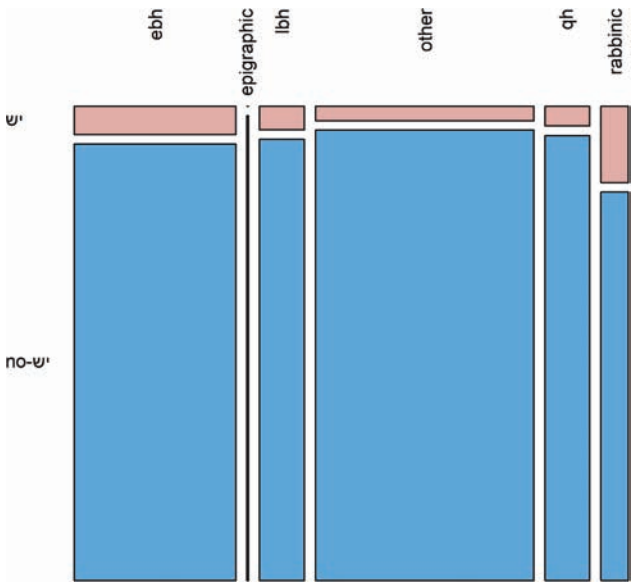

FIGURE 4.21 *Association of language phase and clause type for clauses with and without* יש

Each row of the dataset contains information about one clause. The variables in the dataset are identical to those in the dataset used in section 4.2, although the dependent variable has a different meaning. In the dataset, the output variable cl_type (clause type) has two values. It has the value "jc" if the clause contains the particle יש and "no_jc" if the clause does not contain this particle. The notebooks with which the data are extracted and the resulting csv files can be found on GitHub.[35]

4.3.5. *Data exploration*

First the variables studied throughout the Syntactic Variation project are visualized in relation to the presence or absence of יש, namely, language phase, main and subordinate clauses, discourse, and genre. Also, the count variable clause length is taken into account.

Figure 4.21 shows the distribution of clauses with and without יש in different languages phases.

35 https://github.com/MartijnNaaijer/phdthesis/tree/master/Ch4_Expressions_of_to_be/jc_nojc. The notebook JC_noJC_bib.ipynb preprocesses the biblical data, the file JC_noJC_xbib.ipynb preprocesses the extrabiblical data. The resulting csv-files are jc_nojc_bib.csv and jc_nojc_xbib.csv.

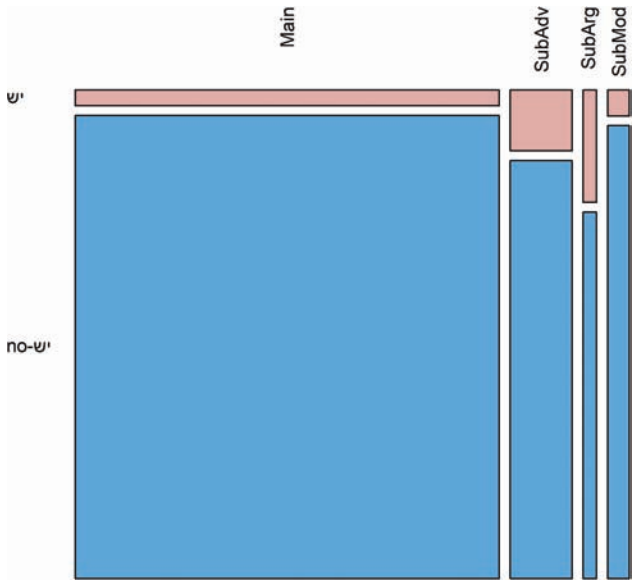

FIGURE 4.22 *Association of main and subordinate clauses and clause type for clauses with and without* יש

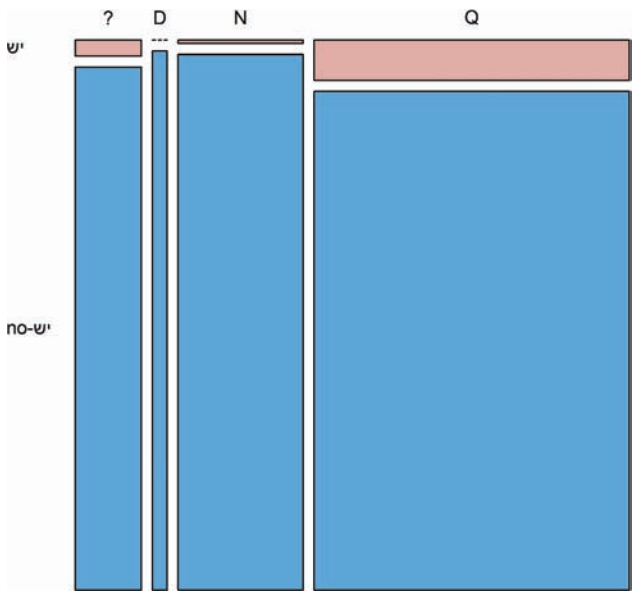

FIGURE 4.23 *Association of discourse and clause type for clauses with and without* יש

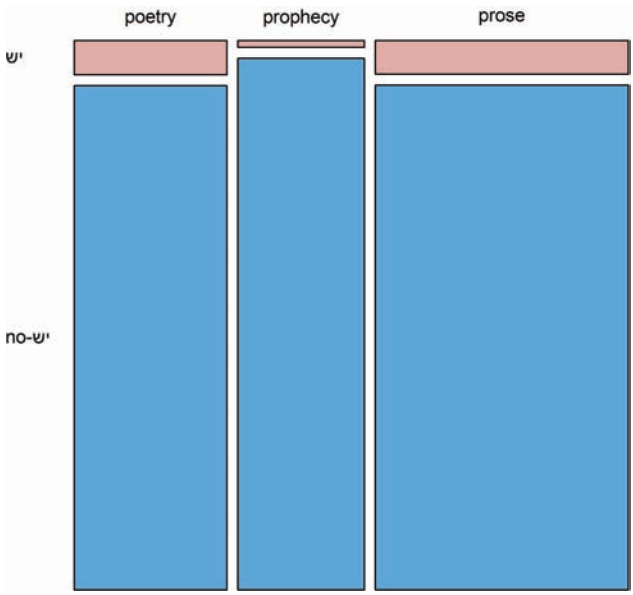

FIGURE 4.24 *Association of genre and clause type for clauses with and without* יש

In section 4.3.1, it was indicated that in the dataset there are 19 times as many clauses without יש than with it. As expected, the mosaic plot shows that in all the environments clauses without יש are more abundant than those with the particle, but some interesting patterns of variation are visible already.

Going from EBH via LBH to QH, there is a slightly decreasing trend in the attestation of יש, but the Rabbinic texts have a far higher attestation of the particle than any other purported language phase. Pirqe Avot and Shirata contain 5 and 8 clauses respectively with the particle and 30 and 31 clauses without it. So, in these Rabbinic texts the ratio of clauses with and without the particle is about 1:4 to 1:8, whereas in the whole corpus it is about 1:19 on average.

Figure 4.22 shows the distribution of clauses with and without יש among main and subordinate clauses.

This figure shows that there is an overrepresentation of the particle יש in subordinate clauses, especially adverbial clauses and clauses that are the argument of another clause.

Figure 4.23 shows the distribution in the different discourse environments, of which quoted speech (Q) and narrative (N) are the most important ones.

This contrast between Q and N is much stronger than between the levels in the other variables under consideration. In N clauses, יש is nearly absent.

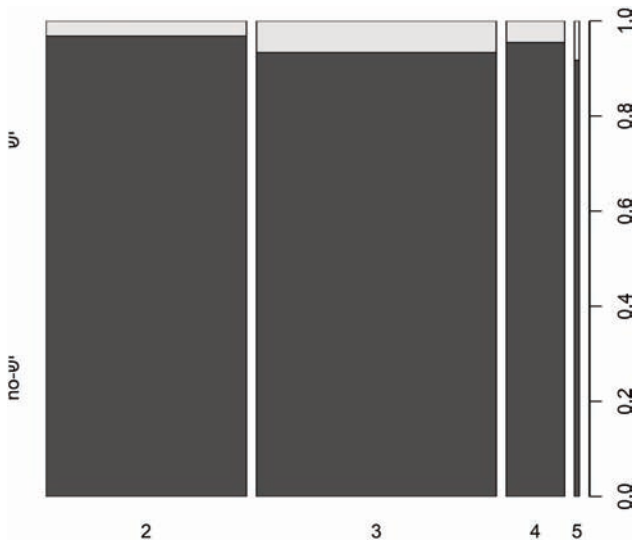

FIGURE 4.25 *Association of clause length and clause type of clauses with and without* יש

Figure 4.24 shows the distribution of יש in the genres of prose, poetry and prophecy.

יש can be found in poetry and prose in more or less equal concentrations, but it is rare in prophetic books.

Figure 4.25 shows the relationship between clause length and the presence/absence of יש.

Clause length is represented here by the number of phrases in a clause and all clauses contain both a subject and a predicate complement, so the shortest possible length is two phrases. To make a fair comparison possible, the particle יש itself is excluded from the counting. The relationship between clause type and clause length looks less clear than in figure 4.9, in which the the presence and absence of היה was visualized in relation to the clause length. In the analysis of היה, there was a pattern in which longer clauses more often contain היה, which is less consistent here.

This exploration suggests that the variable discourse is the strongest predictor for the presence of יש of the main variables of the Syntactic Variation project. There is a strong contrast between the main levels (Q and N) of this variable, and it is likely that this contrast accounts for at least part of the variation in the other variables, especially in the case of the variation in the variable language phase. There is a relatively high concentration of יש clauses in the Rabbinic texts, and most of these are Q clauses.

4.3.6. *Results*

Random Forest

In this section, the results of the Random Forest analysis are discussed first, and then the results of the analysis using XGBoost. The approach used with both algorithms is more or less identical, and the same data are used. The dataset is split randomly in training and test sets, a Random Forest model is trained on the training set and predictions are made on the test set. This procedure is repeated 5 times, each time with a different test set (five-fold cross validation). With the help of the variable importance function we get an impression of which independent variables play an important role in the model. The models are evaluated using the prediction accuracy, the C score and the Receiver Operating Characteristic (ROC) curve.

In the dataset under consideration, the output variable consists of two classes of unequal frequency: יש clauses are rare in comparison with clauses without the particle. In the case of such imbalance, many algorithms tend to classify individual cases from the test set as belonging to the most frequent class. In general, the largest class has the most internal variation in the predictors because there are many samples, so it is difficult to find features that distinguish the samples in the small class from the large class. In the case of the יש dataset, the largest class, clauses without יש, makes up about 95 % of the dataset, so if all samples are classified as "no_jc", an overall accuracy of 95 % is reached, but there is 0 % accuracy for the samples in the smallest class "jc", which means that we know nothing about the difference between the two classes. There are several ways to prevent the model from predicting all observations in the test set as belonging to the largest class of the training set. One of them is to use oversampling (also known as up-sampling). Oversampling is a technique with which the class imbalance is restored by resampling from the smallest class so that both classes have the same size.[36] It is important that oversampling is done after the data are split into train and test data, because if it is done before the split, individual observations can end up in both the training set and test sets, after which the datasets are not independent of each other.

36 Various ways of dealing with class imbalance are discussed by Kuhn and Johnson (2013: 427). I have decided to use a technique known as naïve or random oversampling, which is done by randomly sampling data from the minority class(es) until all classes have equal frequencies. With this technique one avoids the creation of synthetic data, such as is the case with techniques like Synthetic Minority Oversampling Technique (SMOTE).

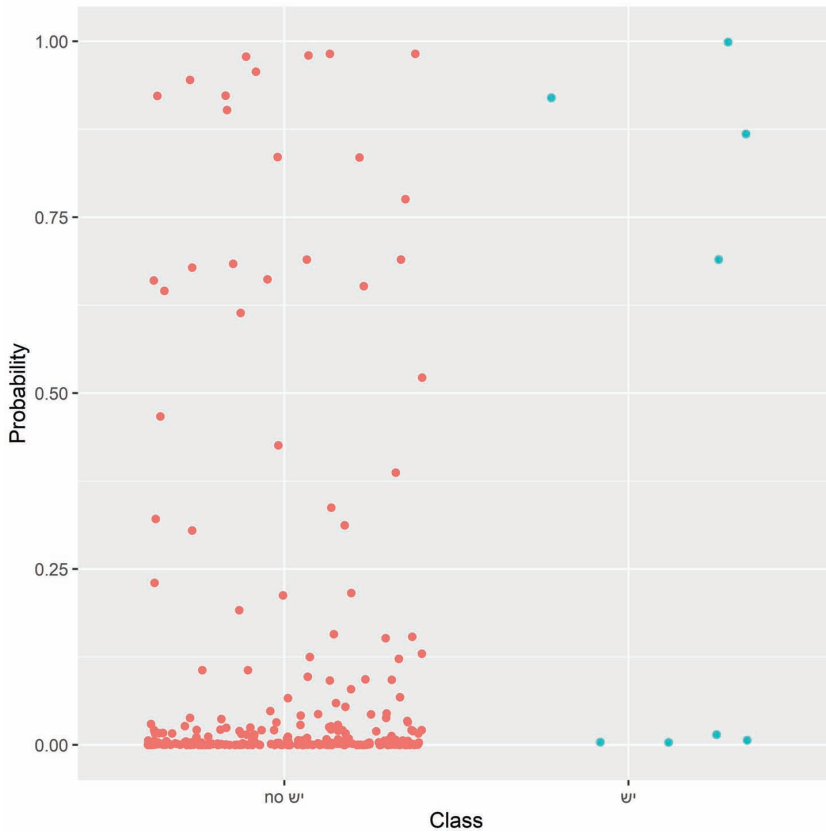

FIGURE 4.26 *Probability of clauses belonging to the classes no-*יש *and* יש

A Random Forest model is made with the datasets described in section 4.3.4. The R-script can be found on GitHub.[37]

After training, a prediction is made for each clause in the test set. The output of the prediction is a score which indicates the probability of a specific observation in the test set belonging to class "jc". This is shown in figure 4.26.

On the left side of the figure, in red, the predicted probabilities of the clauses without יש are plotted, and on the right side, in green, those with the particle can be seen. Each dot represents one clause. By default, probabilities higher than 0.5 are classified as clauses with יש, whereas values lower than 0.5 are classified as clauses

37 https://github.com/MartijnNaaijer/phdthesis/blob/master/Ch4_Expressions_of_to_be/jc_nojc/jc_nojc_rf.R.

without יש. 0.5 is called the cut-off value. A comparison with the true values gives an impression of the strength of the model. In other words, if a sample contains 8 clauses with the particle and a model predicts for each of these clauses that it contains יש, the model seems to do a good job. On the other hand, if the same model has a low accuracy on the other class, the clauses without יש, one should try to improve the model, or change the cut-off value. In practice, it is rarely possible to achieve a 100% accuracy on all classes. If the cut-off is increased, there will be an increase of correct classifications of clauses without יש, but there will be a decrease of the number of correct classifications of clauses containing יש. If the cut-off is decreased, the effect is reversed. So, it is easy to reach a high prediction accuracy on one of the classes in the output variable, but there is always a trade off: increased accuracy on one class generally leads to decreased accuracy on the other class in the case of binary classification.

In the optimization process of the model, I have given both classes equal weight. There is no good reason why the correct classification of one of the classes is more important than the other, so I have decided to maximize the average of the accuracy of both classes.

ROC-curve

The model is evaluated with the C-index, which is derived from the Receiver Operating Characteristic (ROC) curve.[38] Figure 4.27 (see next page) shows the result of the five-fold cross validation in a ROC curve. In the plot for each fold of the validation, the sensitivity is plotted against the specificity of the model. Sensitivity and specitivity are used often to evaluate models with two outcome classes (as יש vs no-יש in this research). The sensitivity is what is called in medical research the fraction of the true positives: this is the group of people with a disease who really have the disease. The specificity is the group of true negatives, these are people who do not have the disease and have a negative, correct outcome of the test. The ROC curve of a perfect model starts in the lower left corner of the plot, moves straight up to the upper left corner and then goes to the right. The curve of a less perfect model moves from the lower left corner to the upper right corner somewhere between the upper left corner and the diagonal of the plot, which signifies a model without any predictive value. A summarization of the model in one number can be given by the C-index (or Concordance statistic). This statistic is the area under the ROC curve. Both axes range

38 The ROC curve is a way to evaluate a model visually. See Appendix C for more explanation.

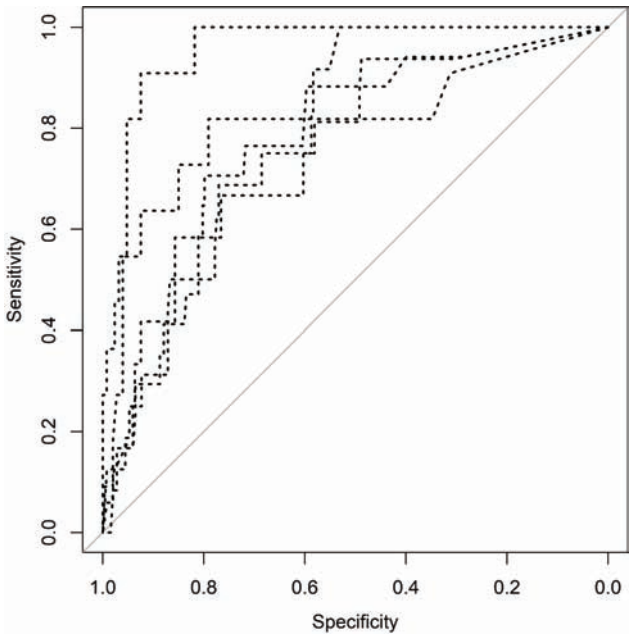

FIGURE 4.27 ROC *curves of five-fold cross validation of* יש *vs no-*יש *using* RF

from 0 to 1, so the perfect model has a C-index of 1 and a non-informative model has a score of 0.5. In this case, predictions are not better than a random guess. The average C-index of the 5 folds is 0.81,[39] which is good (see for instance Tagliamonte and Baayen 2012: 19).

Figure 4.28 shows the average accuracies of the predictions on each of the five test sets. The error bar indicates the confidence intervals of the mean results on the five test sets. In the training set, oversampling was used to balance the classes. This is not done in the test set. On the one hand, this would introduce a new source of randomness, and on the other hand the error bars show clearly the higher uncertainty of the predictions in the יש-class. On the five test sets both classes are predicted correctly between 70% and 80%, but it should be noted that the יש-class is much smaller than the class of clauses without יש.

Other important evaluation metrics for prediction models are precision and recall. These metrics start from the idea that one of the two values is positive and the other

39 The C-indices of the separate folds are 0.87, 0.84, 0.75, 0.83, and 0.76.

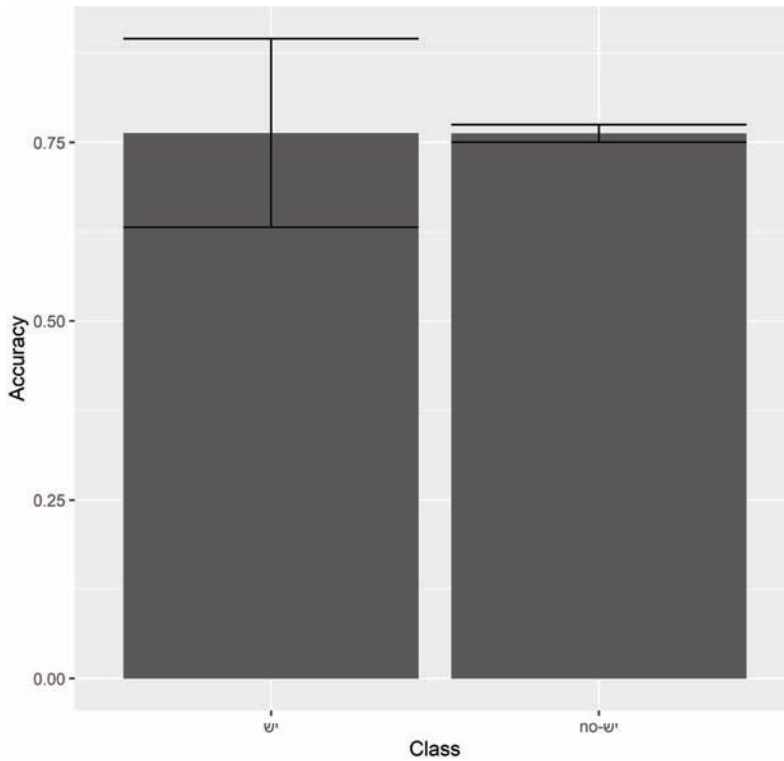

FIGURE 4.28 *Average accuracy of predictions using* RF *with 5-fold* CV *for clauses with and without* יש

is negative. These metrics are important, especially if the classes are imbalanced. Precision is the number of true positives divided by the sum of true positives and false negatives, and ranges between 0 and 1. Often the negative class is much larger than the positive class, and if a small amount of true negatives are classified as positive (false positives), there can easily be more false positives than true positives. This is also the case here. If we say that clauses with יש are positive and clauses without יש are negative, it is clear that the average precision of 0.14 is relatively low, but it is substantially higher than 5%, which is the expected precision in the case of a random classification on the basis of the proportions in the training set.

Recall also ranges between 0 and 1, and is defined as the true positives divided by the sum of the true positives and false negatives. This is equal to the accuracy of the יש-class, which is 0.75.

Making accurate predictions on data is important, but here it is not the only thing that counts. It is clear that the model does an overall good job in making predictions

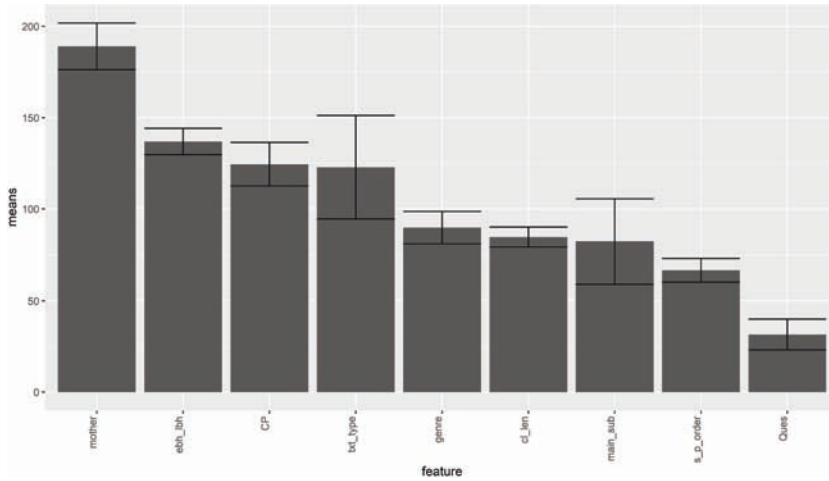

FIGURE 4.29 *Variable importance in the Random Forest model*

on unseen data. Now it is time to look at the role of the individual predictors in the model. Figure 4.29 is the variable importance plot of the model. It shows the average values of the five folds with confidence intervals. In the final model, variables are removed that did not increase the prediction accuracy.

Of the four main variables of this research, language phase discourse type are the most important in the model, although there is also some variation in importance of these variables between the folds. The exploration already showed that יש occurs considerably more in Q than in N, and this is an important factor in the model. There seems to be a certain influence of the language phase on the use of יש, this will be specified further in the analysis using XGBoost.

Slightly less important in this model are the variables genre and whether a clause is a main or subordinate clause. In the mosaic plot in the exploration section, it already became clear that this overrepresentation of יש is strongest in clauses that are the argument of another clause, and most of these clauses containing יש are object clauses.[40]

40 Some examples are: Gen 24:23, הגידי נא לי היש בית־אביך מקום לנו ללין, "Tell me, is there room in your father's house for us to spend the night?"; Gen 42:1, וירא יעקב כי יש־שבר במצרים, "And Jacob saw that there was grain in Egypt"; Gen 42:2, ויאמר הנה שמעתי כי יש־שבר במצרים, "He said, see, I have heard that there is grain in Egypt"; 2 Kgs 4:2, הגידי לי מה־יש־לכי, "Tell me, what do you have".

These two factors combined give the impression that the function of the particle יש is to give emphasis to the clause, as indicated by Muraoka. The weakest predictor in the model is the presence of a question phrase in the clause. Various clauses contain both the particle יש and a question phrase.[41] An explanation for this could be that יש puts extra emphasis on the question that is asked.

Also, in these clauses one gets the impression that there is a certain emphasis. In Num 13:20, יש creates a contrast with אין and in Gen 43:7 and 44:19 יש seems to have an intensifying effect. Clauses with more phrases than subject and predicate tend to contain the particle יש relatively more often, which is also reflected in the model in the variable clause length. An example of this is, of course, a clause introduced with the question-ה, but there is a more general trend than only in the case of question particles.

The strongest predictor is "mother". This is the verbal tense of the clause on which the clause under consideration is syntactically dependent. One can see this as a way to describe how the clause is embedded in its environment.

Figure 4.30 (see next page) shows the association of the values of the variables clause type and mother. The figure shows that clauses without the particle יש are more often associated with a mother clause, which is a verbless clause or a clause containing a wayyiqtol. Clauses with יש are associated more with mother clauses with other verb forms. A complication with this feature is that several values of this variable are associated with values of other variables. For instance, the discourse variable has as main values N and Q, and it is well known that in a N environment other verb tenses are more predominant than in Q. In Q the wayyiqtol is more or less absent, while the imperative and the yiqtol occur frequently.

This dependence of one predictor on another is a problem in various models, but in general the Random Forest model is less sensitive to multicollinearity. If the mother variable is removed from the model the accuracy of the predictions of the value יש decreases, so the presence of the variable mother cannot be explained completely by other predictors. The same is true the other way around. If the discourse type is removed, the accuracy of the prediction of יש also decreases, therefore it is kept in the model.

---

41 Next to the example in Gen 42:23: Gen 43:7, לאמר העוד אביכם חי היש לכם אח, "... saying, 'Is your father still alive? Do you have another brother?'"; Gen 44:19, לאמר היש־לכם אב או־אח, "... saying, Do you have a father or a brother?"; Num 13:20, היש־בה עץ אם־אין, "And whether there are trees in it or not"; Job 6:30, היש־בלשוני עולה, "Is there any wrong on my tongue?"; Job 25:3, היש מספר לגדודיו, "Is there any number to his armies?"; Job 38:28, היש־למטר אב, "Has the rain a father?".

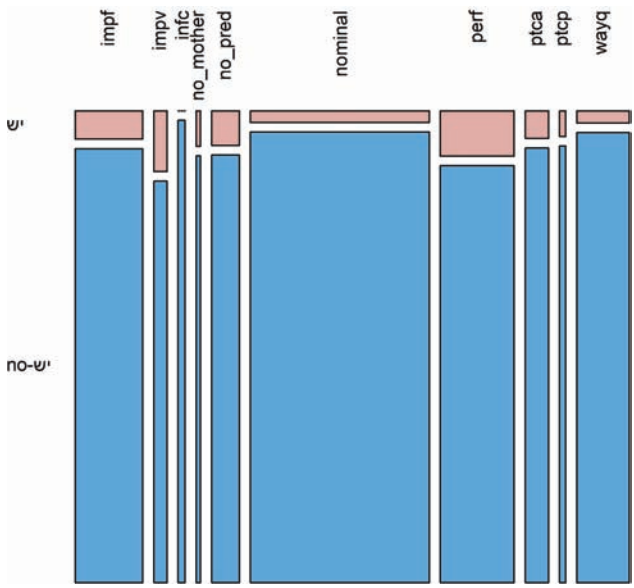

FIGURE 4.30 *Association between clause type and tense of the mother of the clause*

Extreme Gradient Boosting

The analysis is repeated using XGBoost. The goal of doing the analysis with a different algorithm is to find out on the one hand if it is possible to achieve a higher prediction accuracy and, on the other hand, to test the robustness of the first analysis with Random Forest.

In the implementation of the package xgboost, the algorithm needs the data as a sparse matrix. This means that categorical variables need to be split in several separate binary variables, one for each level of all the original categorical values. In the present analysis, the data are the same as in the Random Forest analysis, but I have not reduced the number of variables to make the data easier to interpret. The overall prediction accuracy is similar to the analysis using Random Forest.

ROC-curves

The average area under the curve of the ROC plots is 0.83.[42] This is a slightly better situation than the one in the random forest analysis, but overall, the ROC curves (figure 4.31, next page) are similar to those in the Random Forest analysis.

42 The separate values are 0.811, 0.76, 0.87, 0.81, and 0.90.

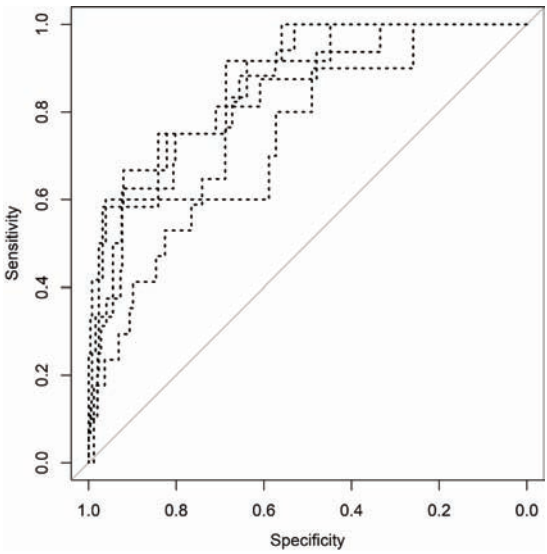

FIGURE 4.31 ROC *curves of five-fold cross validation of* יש *vs no-*יש *using XGBoost*

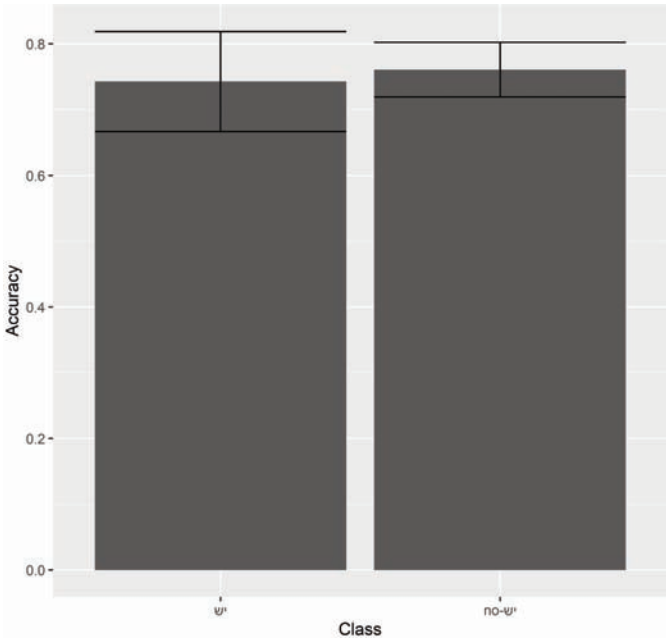

FIGURE 4.32 *Average accuracy of predictions using XGBoost with 5-fold* CV *for clauses with and without* יש

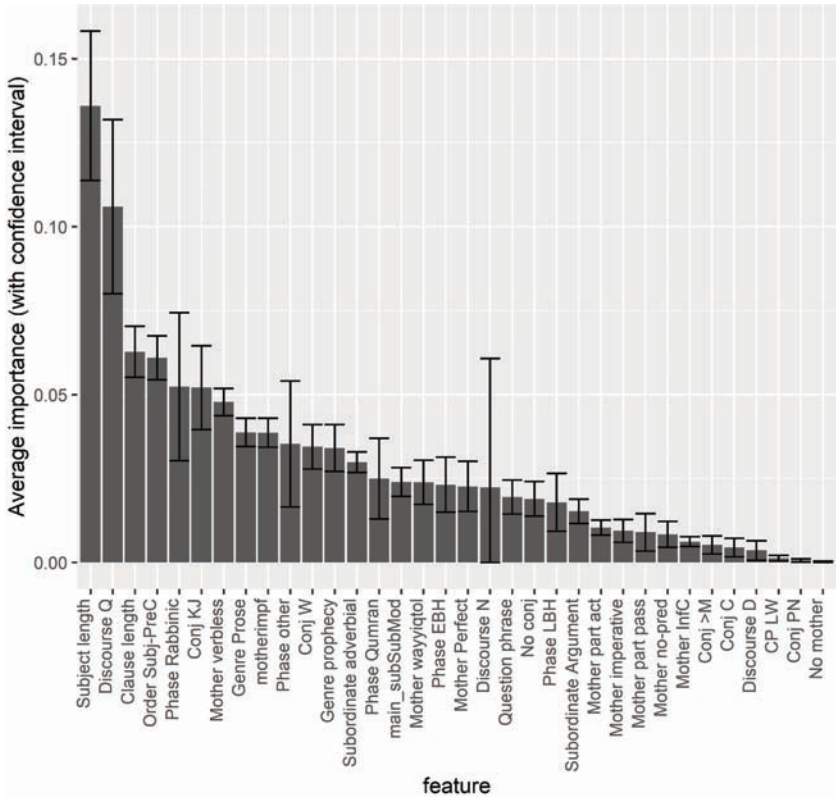

FIGURE 4.33 *Variable importance in the XGBoost model*

Also, the accuracies of the predictions on the test set are more or less equal to those in the Random Forest analysis; see figure 4.32.

An interesting insight is given by the variable importance plot. In the XGBoost analysis, the number of predictors has not been reduced as in the Random Forest. In the plot in figure 4.33, all the variables in the model can be observed.

Every level of every categorical variable in the original dataset becomes a separate variable in the XGBoost model. This makes it easier to interpret the influence of the separate levels. The most important one related to our main variables is discourse type Q. From the exploration it was clear already that there is an overrepresentation of יש in Q clauses. Also visible is the high concentration of יש in the Rabbinic texts. Two of the most important variables are numeric. These are the lengths of the subject (in words) and the length of the clause (in phrases). The importance of the numeric variables in this analysis is due partly to the fact that the categorical variables are split

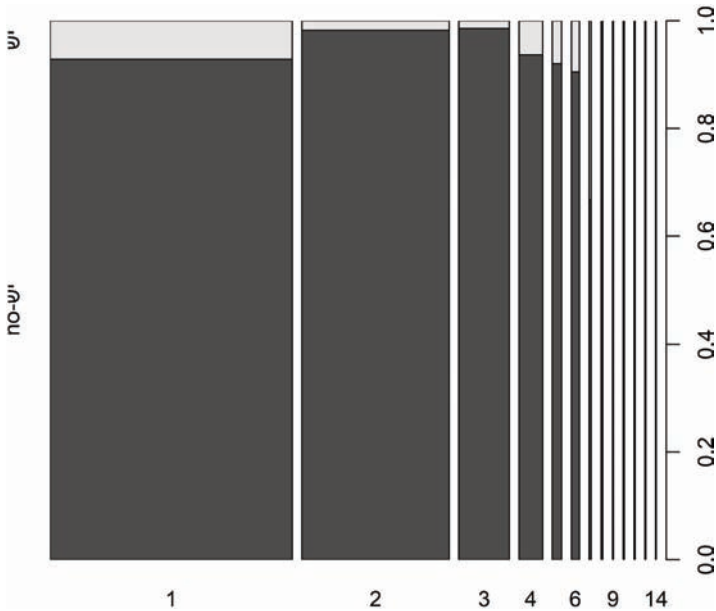

FIGURE 4.34 *Association of subject length and clause type of clauses with and without* יש

into binary variables, by which the present numeric variables grow in importance. In the Random Forest model, they were dropped, because their inclusion did not improve the prediction accuracy.

Figure 4.34 shows the relationship between subject length and clause type. The figure shows that clauses with יש have a preference for short subjects with one word and subjects with subjects longer than 3 words.

All in all, the analysis with XGBoost does not reveal totally new things, neither does it improve the prediction accuracy relative to the Random Forest model, but it confirms the results of it and shows slightly clearer which levels of which variables are important for separating clauses with and without the particle יש.

Discussion of the results

The results of this analysis cohere well with the idea that the particle יש puts emphasis on the clause, as argued by Muraoka. The particle is relatively rare, but there is a clear association of the particle with Q clauses, and object clauses, which often follow verbs of saying or perception. Also, it occurs often in Rabbinic texts, whose language is generally associated with vernacular speech. Although all these observations do not necessarily imply emphasis, they fit the idea well. There does not seem to be much

empirical evidence for the hypothesis that there is an increased use of יש in the course of the biblical period, as Schüle suggests. If there is a tendency visible, it is that יש is used less in the constructions under consideration, going from EBH to LBH.

There seems to be a relationship between subject length and clause length and presence/absence of the particle יש, although this relationship between clause length and clause type is not linear. Van Hecke showed already that the length of the predicate complement plays a role in the structure of the clause, although in a different way than is shown here. This could mean, that יש is used to give structure to the clause.

4.4. The tripartite verbless clause

4.4.1. *Problem and research question*

An intensely discussed topic in the literature on the verbless clause is the grammatical status of the so-called tripartite verbless clause. The tripartite verbless clause has the following structure:

NonVerbalPhrase - PPrP - NonVerbalPhrase

or

NonVerbalPhrase - NonVerbalPhrase - PPrP

An example of the first type is Deut 12:23, כי הדם הוא הנפשׁ, "For the blood is the breath", and an example of the second type is Jer 10:3, כי־חקות העמים הבל הוא, "For the customs of the peoples are vanity".

On the one hand, the clause can be interpreted as a clause in which the pronoun functions as a copula. In this interpretation, the clause is often called a tripartite clause. In this research, I will also use this expression, without necessarily indicating that the pronoun functions as copula. According to others, the pronoun is the subject of a bipartite clause, or the pronoun gives emphasis to an extraposed subject. Finally, it has been hypothesized that the pronoun functions as a particle in this kind of constructions.

In most studies, the validity of the "copular theory" is investigated by giving a number of examples of the structure that support a specific interpretation. In this research, I choose a different approach. Just like in the previous sections, all cases of the

tripartite clause and their environments are contrasted with those of the "ordinary" bipartite clauses occurring in the Hebrew Bible. The research question of this section is:

> On the basis of the structure and environment of tripartite verbless clauses, which variables play a role in distinguishing between bipartite and tripartite clauses?

On the basis of the results, an evaluation will be made of the function of the personal pronoun in the tripartite verbless clause. Which of the different approaches accords most with the data? This question will be investigated with the core variables of the Syntactic Variation project using Random Forest and XGBoost.

4.4.2. *Review*

First, the works of a number of scholars who interpret the pronoun as a copula in BH are discussed, thereafter the alternative points of view will be studied. Also, the evidence from Syriac will be taken into account, because this comparative evidence is used by some scholars to show that BH uses the pronoun as a copula.

The pronoun as copula

Originally the Semitic languages did not have a copula, but developed this in two ways, according to Albrecht (1888). The first way is the development of the pronoun as separator (Trennungswort) between subject and predicate, especially if both clause members are determined (Albrecht 1888: 250). The second way is the development of the verb היה as copula, which first had the meaning "to exist", and was used in the later parts of the Hebrew Bible as copula, according to Albrecht (1888: 252–253) The evidence that the pronoun is used as a copula in Hebrew is derived by Albrecht from Ethiopic, in which the pronoun often functions as a copula.[43] Another argument for the existence of the copula in Hebrew is that the resumption of a non-third person subject by a third person pronoun is rare in Hebrew (Albrecht 1888: 251–252).

43 Albrecht borrows the argument from Ewald (1855, § 297b).

According to Brockelmann (1956 §30a) the pronoun was used as a real copula in BH, which became possible after its original meaning was lost. He critizes Driver, stating that he misunderstands the meaning of the pronoun.[44] Brockelmann gives a number of examples of the pronominal copula and comments that it may occur in various constructions, not only when both subject and predicate are determined. It is also used in clauses with an AdjP as predicate,[45] in clauses starting with the predicate complement,[46] in clauses in which the subject is a first- or second-person pronoun,[47] and even in clauses with a verbal predicate.[48] It is this variation in the use of the pronoun that Brockelmann sees as evidence for the pronominal copula. In later publications, some scholars opted for this interpretation of the pronoun, generally using arguments that are similar to those of Albrecht and Brockelmann.

The pronoun does not function as a copula

In Appendix V of "A Treatise on the Use of the Tenses in Hebrew and Some Other Syntactical Questions", Driver discusses the use of the casus pendens in Biblical Hebrew. The casus pendens plays an important role in avoiding sentences in Hebrew that are too complicated, by giving prominence to the subject and "easing" the clause by the use of the pronoun (Driver 1892a: 265). From this perspective, Driver investigates whether the pronoun can express the copula. He discusses a case in which the pronoun follows the predicate in Gen 34:21 (footnote 111).

This could also be expressed without the pronoun, but the way it is expressed in the biblical verse makes the sentence "less cumbrous and less abrupt", and this advantage becomes more important when the subject contains a long relative clause[49] (Driver 1892a: 268). In other words, the use of the pronoun has a stylistic goal and it does not function as a copula. In the case of Gen 34:21, the subject was resumed without emphasis, but there is emphasis in the case where the pronoun stands before the predicate. Driver (1892a: 269) distinguishes between clauses with a definite and indefinite predicate. If the predicate is definite, the subject is identical to the predicate,

---

44 According to Muraoka (1983: 67 n. 1), this is a misrepresentation of Driver's position.

45 Gen 34:21, האנשים האלה שלמים הם אתנו, "These people are at peace with us".

46 Lam 1:18, צדיק הוא יהוה, "Righteous is YHWH"; Song 6:9, אחת היא יונתי, "One is my dove".

47 Isa 52:6, כי־אני־הוא המדבר, "That it is I who speak"; Ps 44:5, אתה־הוא מלכי, "You are my king".

48 1 Kgs 18:24, והיה האלהים אשר־יענה באש הוא האלהים, "The god who answers by fire is indeed God".

49 This reminds one of the results of section 4.3. Clauses with a longer subject tend to contain the particle יש more often.

and if the predicate is indefinite it defines the class of the subject. In many cases, it is clear to Driver that the use of the pronoun is unmistakably emphatic, both with definite and indefinite predicate.[50]

For Driver, a strong argument against the use of the pronoun as copula is the parallelism in various verses in which the pronoun is used first in a participle clause and then again with a finite verb (Deut 9:3, 31:3 and 8, Josh 22:22,[51] Driver 1892a: 270 n2). He also states that in the course of time the pronoun "lost its distinctive force, and became little more than the copula", although this development is not really visible in EBH and LBH (Driver 1892a: 270 n. 4).

In a separate section, Driver discusses cases in which the resumptive pronoun does not agree in person with the fronted subject when this is also a personal pronoun.[52] Even in these cases Driver does not consider the pronoun to function as a copula. Driver refers to Roorda and Delitsch, who suggest that the meaning of the clause אתה־הוא האלהים in Isa 37:16 is: "Thou, he (and none else) art my king" or the pronoun could anticipate the predicate, according to Ewald: "Thou art he-my king" (Driver 1892a: 271).

Muraoka has made a contribution to the study of the tripartite verbless clause in several works.[53] In "Emphatic Words and Structures in Biblical Hebrew" (1983), he implicitly criticizes the idea that the pronoun can function as a copula with a quantitative argument: "A convincing case for the existence of pure copula in Biblical Hebrew could be made if one could adduce a meaningful number of examples" (Muraoka 1983: 68).

Muraoka distinguishes three types of constructions in which the copular pronoun plays a role. The first of these are clauses with the structure S-C-P (Subject - Copula - Predicate) (Muraoka 1983: 72). According to Muraoka this pattern is often selective-exclusive. The extraposed subject is contrasted with alternatives.[54]

---

50 He gives the examples Deut 4:35, 39, similar in 7:9, 1 Kgs 18:39, יהוה הוא האלהים, "YHWH, he is God"; Deut 3:22, כי יהוה אלהיכם הוא הנלחם לכם, "For YHWH your God, is fighting for you"; Deut 9:3, כי יהוה אלהיך הוא־העבר לפניך; "That YHWH your God, he crosses before you"; Deut 31:6 and 8, כי יהוה אלהיך הוא ההלך עמך, "For YHWH your God, he is the one walking with you".

51 For instance, the case in Deut 9:3 is כי יהוה אלהיך הוא־העבר לפניך אש אכלה הוא ישמידם והוא יכניעם לפניך, "That the LORD your God is the one who crosses over before you is a consuming fire; he will defeat them and he will subdue them before you".

52 For instance, Ps 44:5, אתה־הוא מלכי, "You are my king"; Isa 37:16, אתה־הוא האלהים, "You are my God".

53 The main contributions are in JM § 154 i–m, Muraoka (1983 and 1999).

54 For instance, Deut 4:35, כי יהוה הוא האלהים אין עוד מלבדו, "That YHWH is God, there is no one

The second type is S-P-C (Subject-Predicate-Copula), which Muraoka calls descriptive (Muraoka 1983: 75), in which the predicate is often an adjective or a prepositional phrase. Similar to the S-C-P clauses the accent pattern is often $S^{\vee}$P-C or $S^{\vee}P^{\vee}$C, which is a confirmation of the isolation and emphasis on the extraposed subject and the predicate.

The third and last type is P-C-S (Predicate-Copula-Subject). The accent pattern is always P-$C^{\vee}$S. P-C is a unit in this kind of clauses, which is isolated from the subject.

All in all, Muraoka rejects the idea that the pronoun can function as a copula in Biblical Hebrew and also in Semitic in general. He is also sceptical of theories in which it functions to give structure to the clause, as Driver argues. For Muraoka (1983), the pronoun has the function of giving emphasis or prominence to the preceding NP.[55]

A middle position

A middle position is defended by Holmstedt and Jones (2014, abbreviated as HJ). While most linguists choose for or against the existence of the pronominal copula, HJ argue, like Sappan (1981), that in a number of cases of the potential candidates the pronoun functions as copula and in the other cases it functions as resumptive pronoun. HJ criticize other scholars by stating that they did not make good "contextual sense with all the examples" (HJ 55). The first example they mention of a clearly resumptive case is Gen 34:23 מקנהם וקנינם וכל־בהמתם הלוא לנו הם, "Will not their livestock, their property, and all their animals be ours?", in which the question-ה clearly introduces a clause (HJ 56). Most candidate clauses are cases of left (front) dislocation, which presents a Topic-Focus structure (HJ 57). The "dislocated constituent is the Topic, orienting the reader either to which entity (among multiple discourse possibilities) the following clause adds information or to scene-setting information (e.g., place, time)."

HJ argue that most cases of the tripartite clause in BH are cases of left or right dislocation, but in some cases this analysis would be "infelicitous", and in these cases an analysis of the pronoun as copula is more appropriate (HJ 58). HJ give four different arguments for the existence of the copular pronoun in BH.

The first kind of clause in which resumption does not fit well is when there is no agreement between the fronted element and the copula (HJ 59), for example, Ps 44:5 אתה־הוא מלכי אלהים, "You are my king, God". In this case, resumption is impossible

beside him".

55 See also the excursus on the pronominal copula in Syriac below.

because there is no agreement between the second person אתה and the third person הוא. An alternative translation of the clause would be "You are he, my king, God". In this translation, the clause is interpreted as a case of right dislocation, but the problem is that הוא needs a third person antecedent. This role cannot be fulfilled by מלכי, because it follows the pronoun, and it is also missing implicitly or explicitly in the preceding context of Ps 44:5. From this, HJ conclude that הוא functions as copula in this clause.[56]

A second argument for the existence of the pronominal copula in BH is that in both distribution and structure there are four parallels with the copula היה. The first is that both clause types may occur with a NP, PP, or AdjP predicate, although most cases of the pronominal copula occur with a NP predicate (HJ 59–60, also n. 17). HJ argue that both היה and the pronominal copula indicate the tense of the clause (HJ 61). In the case of היה, this is past and non-past tense and the pronominal copula indicates present tense. The copular clauses form a contrast with the bipartite verbless clause, in which a tense marker is absent. The second parallel is that the negation of the pronominal copula matches the expectation, because it is negated with אין in Gen 44:26, just like the participle is generally negated. Likewise, the היה clause is negated by לא, just as other verbal clauses with perfect or imperfect forms (HJ 61–62). The third parallel is the possibility of occuring with participial clauses (HJ 62). The fourth parallel is that in clauses with היה or a pronominal copula, verb raising can occur. This means that in the presence of certain constituents the normal word order of SV is inverted to VS. This can happen if the clause is initiated by subordinators like כי or למען, in the case of modality, including negators, or if there is a focus constituent, such as a clause which starts with a focused predicate (as in Gen 46:34, HJ 62–63). This phenomenon of V-raising occurs once in a clause with a pronominal copula in Isa 9:14, זקן ונשׂוא־פנים הוא הראש, "The old and honorable is the head".

HJ's last main argument in favor of the idea that the pronoun can function as a copula in BH is based on a historical/comparative analysis of the phenomenon. The way the copula developed in Biblical and post-Biblical Hebrew is similar to how it developed in other Semitic languages (HJ 71–73).

56 In the case of disagreement between the resumed subject and the pronoun, Muraoka argues that the pronoun functions as a particle, instead of as a pronoun (KP: 158).

Excursus: the tripartite verbless clause in Syriac

It is relevant for the study of the pronoun as a copula in BH to have a look at its counterpart in Classical Syriac, because the evidence in other languages is often part of the discussion on the tripartite clause in BH. In "Corpus Linguistics and Textual History" by Van Keulen and Van Peursen, eds. (2006, abbreviated KP), five articles are dedicated to the issue of the tripartite clause in Syriac.[57] Van Peursen explains the three positions scholars have proposed about this issue in "Three Approaches to the Tripartite Nominal Clause". The first position is that of Geoffrey Khan, according to whom the pronoun functions as a copula in Syriac.[58] Khan's argument in favor of the interpretation of the copular pronoun are its common use in Syriac, and the fact that it does not seem to be marked there (KP 157).

The second position is that of Goldenberg. According to him all verbless clauses are derived from a basic clause type P-S. In such a clause, the pronoun is the subject which resumes an extraposed subject (KP 160–161).

The third opinion is that of Muraoka, and according to him the pronoun is an emphatic particle, just like in Hebrew (KP 158). Van Peursen gives two reasons why Muraoka has a strong position (KP 161). First, the pronoun can follow any part of speech, and is not part of some fixed clause structure, as is assumed by others. This makes it easier to give the explanation of the function of the pronoun instead of having to explain every distinct environment. Second, the pronoun occurs in positions where it does not agree with the subject of the clause. This means that the pronoun functions as a particle and not as a coreferential pronoun.

All three positions have their pros and cons, but a main difference between Hebrew and Syriac is that the pronoun in a tripartite construction is far more frequent in Syriac than in Hebrew, which is exactly one of the reasons why Muraoka (1983) rejects the existence of the pronominal copula in Hebrew.

57 On the pronominal copula in the Hebrew book of Kings and its translation in Syriac, see Dyk and Van Keulen (2013: 409–411).

58 This is also the vision of the grammars of Nöldeke, Duval, and Costaz.

4.4.3. *Data preparation, experimental approach and variables*

From the ETCBC database all relevant bipartite and tripartite clauses were extracted. The data and the Text-Fabric Scripts with which the data are prepared can be found on GitHub.[59]

The experimental approach in this section is similar to that in the analysis in section 4.3 on variation between clauses with and without the particle יש. With Random Forest and XGBoost it will be predicted whether a clause is a bipartite or a tripartite clause. From this, I want to find out which environments have a preference for the tripartite clause, as compared to bipartite clauses.

The predictors used in the analysis of the tripartite verbless clause are fewer than those in the sections 4.2 and 4.3. In those analyses, the structure of a clause played an important role, for instance, features related to the length of a clause or the presence of other phrases with specific functions. In the case of the tripartite clause, these are more difficult to describe unambiguously, because the status of the clause itself is ambiguous. For instance, in the case of the following clause in Gen 2:14, והנהר הרביעי הוא פרת, "And the fourth river is the Euphrates", the subject is a NP if it is interpreted as a tripartite clause, but it would be PPrP if the clause is seen as a bipartite clause with resumption. To avoid complications, I have decided only to use the core variables of the Syntactic Variation project.

4.4.4. *Data exploration*

The tripartite verbless clauses in Classical Hebrew form a limited group, I have counted 175 cases in the MT and 26 cases in the extrabiblical texts. As the review has shown, these clauses have been discussed extensively from a qualitative perspective. It is clear that the phenomenon of a clause with a tripartite structure exists, but what can a quantitative interpretation add to its interpretation? A good suggestion was made by Muraoka, who stated that the low frequency of the tripartite clause is a

59 The relevant files are in the folder https://github.com/MartijnNaaijer/phdthesis/blob/master/Ch4_Expressions_of_to_be/bipartite_tripartite. The tripartite clauses can be found in four csv-files: tripartite_bib.csv, tripartite_xbib.csv, tripartite_eppr_bib.csv, tripartite_eppr_xbib.csv. The files with "eppr" contain tripartite clauses in which there is no agreement between the resumed phrase and the pronoun. The bipartite alternatives are extracted from the file hyh_nom_bib.csv and hyh_nom_xbib.csv from the folder https://github.com/MartijnNaaijer/phdthesis/tree/master/Ch4_Expressions_of_to_be/hyh_verbless. The files with "xbib" contain extrabiblical data. These files are merged and analyzed in the file bip_trip_plots_analysis.R.

TABLE 4.5 *Frequencies of bipartite and tripartite clauses in Classical Hebrew*

	Biblical texts	Extrabiblical texts
Bipartite clauses	7203	639
Tripartite clauses	175	26

sign that the pronoun does not function as a real copula. Also, if it does function as a copula, it is likely that it exists next to cases of resumption, because there is little doubt that resumption exists. This is the option that HJ proposes, but this makes the group of potential copula clauses only smaller: in appendix B of HJ, 38 copular cases are mentioned. These low frequencies are a strong sign that at least the pronominal copula has a low productivity. On the other hand, there may be specific environments in which it occurs more often, which could be a sign of language variation or change.

The recipe here is the same as in the other expressions of "to be" in BH. An exploration is made of the bipartite and tripartite verbless clause in relation to the language phase, the discourse type, genre, and whether the clauses are main or subordinate clauses.

As was the case in the analysis of clauses with and without יש, in all environments the tripartite clauses form a small minority, and the contrast between the sizes of the small and large classes is even greater here than in the case of the יש-analysis, as table 4.5 shows.

About only 2.5% of the dataset consists of tripartite verbless clauses.

Figures 4.35–38 show the distribution of bipartite and tripartite clauses throughout the different levels of the core variables of the Syntactic Variation project.

The most striking results are that the tripartite clauses occur relatively more often in subordinate clauses than main clauses, more in Q and D than in N clauses, and, finally, slightly more in prose than in poetry and prophecy. These results cohere well with the opinion that in the tripartite clause in BH the pronoun gives emphasis or prominence to the fronted subject, as one would expect in quoted speech.

Similar to יש clauses, tripartite clauses occur relatively often in subordination (although not in attributive clauses), especially as an argument of another clause. Also, from the language phase plot it becomes clear that the tripartite clause is relatively rare in LBH in comparison with EBH, and occurs most often in the RH texts, and of the three genres tripartite clauses occur a bit more in prose than in the other genres, but the contrast is not strong.

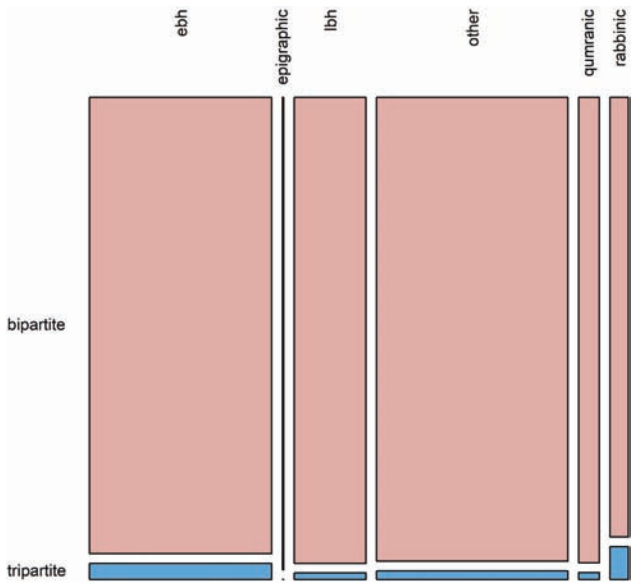

FIGURE 4.35 *Association of language phase and clause type for bipartite and tripartite clauses*

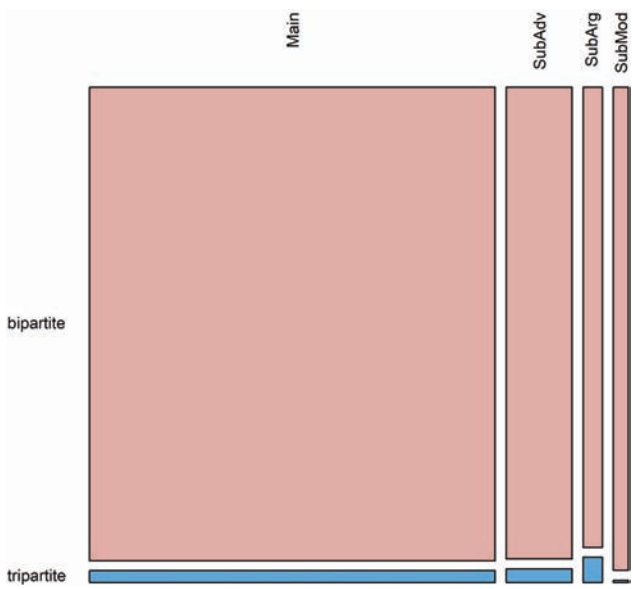

FIGURE 4.36 *Association of main and subordinate clauses and clause type for bipartite and tripartite clauses*

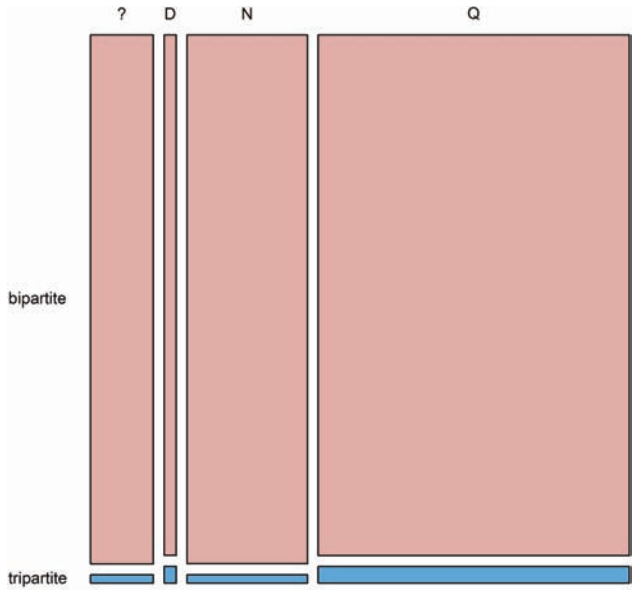

FIGURE 4.37 *Association of discourse type and clause type for bipartite and tripartite clauses*

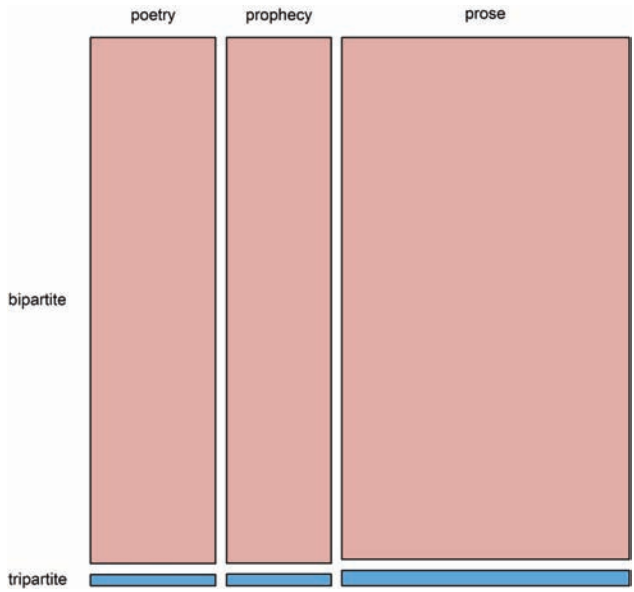

FIGURE 4.38 *Association of genre and clause type for bipartite and tripartite clauses*

4.4.5. *Results*

The algorithms used for this analysis are Random Forest and XGBoost. The dataset consists of the Hebrew bipartite and tripartite verbless clauses in the MT and the texts available in the extrabiblical Text-Fabric package.

The Random Forest model is trained on the data with the clause type (cl_type) as dependent variable. This variable has two possible values: "bipartite" or "tripartite". The predictors are the language phase (ebh_lbh), main or subordinate clause (main_sub), discourse type (txt_type) and genre (genre). The validation of the model is done with five-fold cross validation. Oversampling is used to balance class sizes in the training sets.

Results of the Random Forest

Figure 4.39 shows the ROC plot.

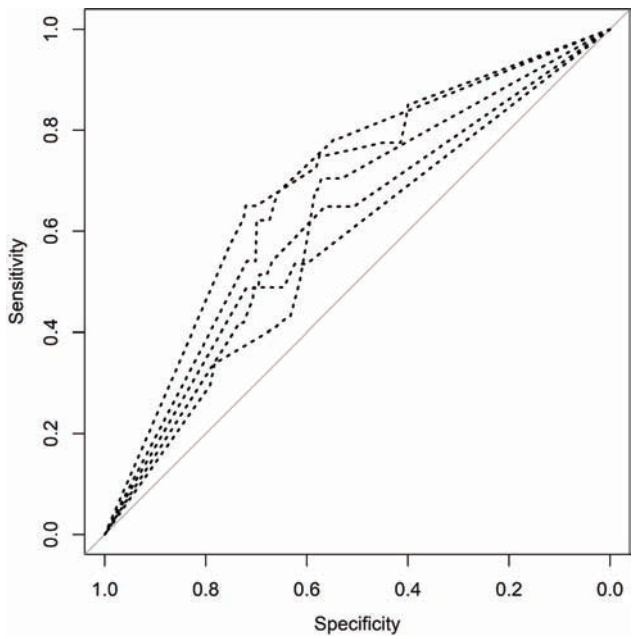

FIGURE 4.39 ROC *curves of five-fold cross validation of bipartite and tripartite clauses using Random Forest*

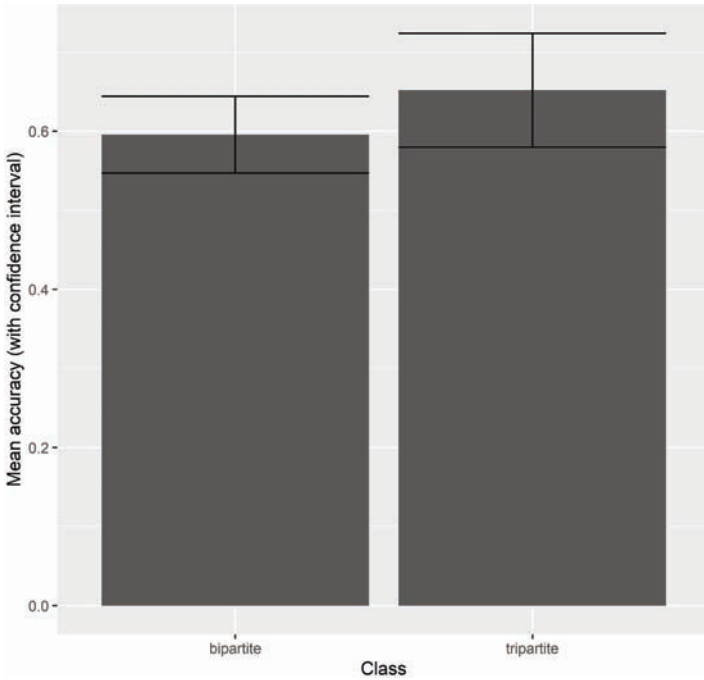

FIGURE 4.40 *Prediction accuracy of the Random Forest model for bipartite and tripartite clauses*

The average C score is 0.65.[60] This C score is lower than in the case of the model in section 4.3. This is not surprising, because the model was trained with fewer predictors.

Figure 4.40 shows the prediction accuracy for bipartite and tripartite clauses with the Random Forest model.

Compared with the results of the analysis of clauses with and without יש, the RF models performed a bit worse on distinguishing bipartite and tripartite clauses. There can be various reasons for this. In the first place, the nature of the data can cause a different performance. If one wants to distinguish two different classes and the predictors have similar values, it is more difficult to distinguish them. In the second place, only the main variables of the Syntactic Variation project have been taken into account. It is likely that this smaller set of features has an important effect on the performance of the model.

60 This is the average of the five folds: 0.6709, 0.632, 0.6333, 0.6272, and 0.6945.

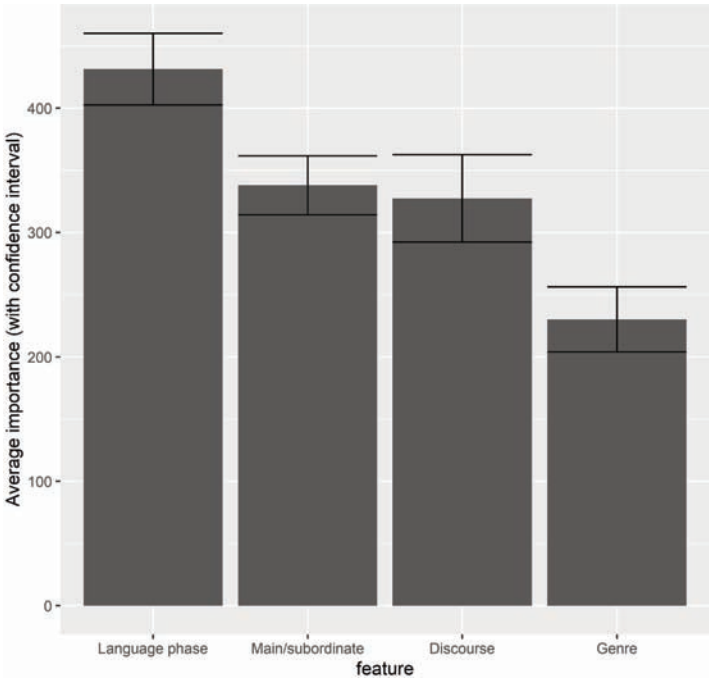

FIGURE 4.41 *Variable importance of the Random Forest model*

Another related cause of different performance of the models can be the number of observations in the smallest class. If the smallest class has only a small number of observations, it is likely that the variation in the data are not representative of the natural use of a certain construction. This is a general problem in the study of ancient languages, for which we have only limited data. Also, it is not only a problem of the smallest class, because the variation in the largest class also suffers from the limitations of the whole corpus, but generally, if one wants to investigate the difference between two classes, the size of the smallest class is an important limiting factor. In the case of the bipartite/tripartite analysis, the smallest class (the tripartite clauses) contains 201 clauses, whereas in the case of the יש-analysis, the smallest class (the clauses containing יש) contains 67 clauses. Therefore, the decreased performance of the models in the bipartite/tripartite analysis is probably not caused by the problem of the limited number of observations in the smallest class.

Figure 4.41 shows the importance of the four variables in the RF model.

Language phase is the most important variable in the model, followed by main and subordinate clauses and discourse type have similar importance scores, especially

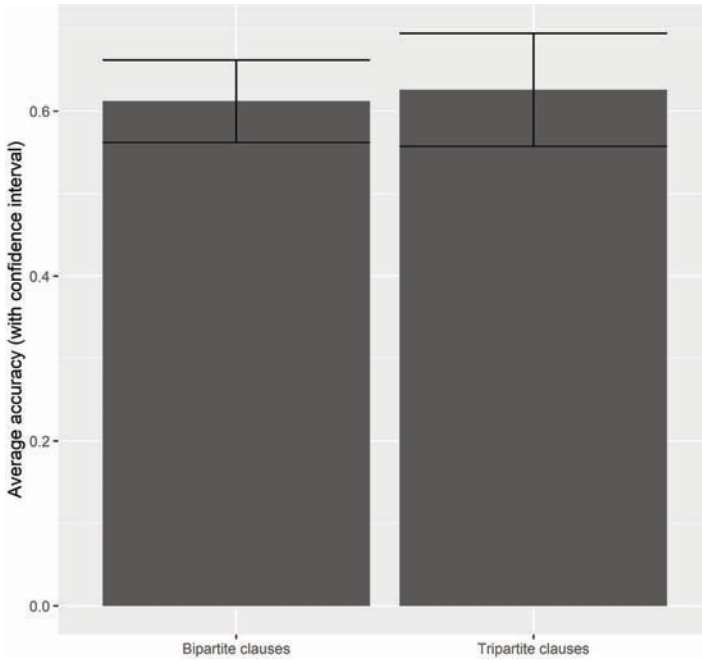

FIGURE 4.42 *Prediction accuracy of the XGBoost model on bipartite and tripartite clauses*

if the confidence intervals are taken into account. Genre is less important, which is as expected, because the mosaic plot of genre showed that tripartite clauses are distributed more or less evenly in the three genres.

Results of the Extreme Gradient Boosting model

The same analysis of bipartite and tripartite clauses was done with XGBoost. Upsampling was used to balance the class sizes of the training sets and five-fold cross validation was applied. Figure 4.42 and 4.43 show the prediction accuracies on bipartite and tripartite classes, and the corresponding ROC-plot.

Each class in the test set is predicted correctly with a probability on average of about 0.60–0.65. Figure 4.43 shows the corresponding ROC-plot. The average C-score is 0.67.[61] This means that the XGBoost model performs slightly worse than the Random Forest model.

Figure 4.44 shows the variable importance plot of this model.

61 This is the average of the five folds: 0.681, 0.7047, 0.6735, 0.6315, and 0.6675.

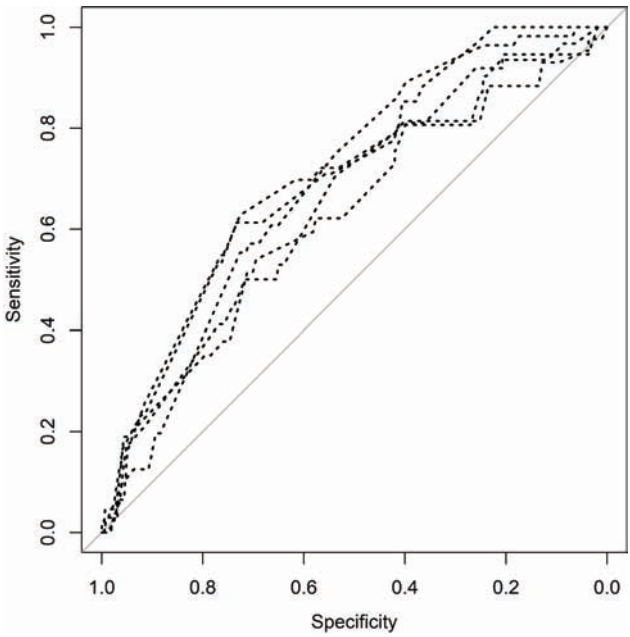

FIGURE 4.43 ROC *curves of five-fold cross validation of bipartite and tripartite clauses using XGBoost*

The most important variable in the model are whether a clause is a Q clause, and two of the levels of main and subordinate clauses, namely whether a clause is an argument clause or a modifier clause. The figure complements the results of the Random Forest model. Tripartite clauses have a preference for the Q environment. The simplest explanation for this is that in these clauses the pronoun emphasizes or gives prominence to the resumed subject. Quoted speech is the natural place for this. There is no obvious reason why the tripartite clause with the pronoun functioning as a copula would occur more in quoted speech than in narrative.

The second most important variable is whether a clause is an argument clause. These verbless clauses have a tripartite structure relatively often. In the Hebrew Bible, these are all object clauses, being the object of the verb ידע, "to know", or a verb of speaking or perception. Generally, there seems to be emphasis on the object clause, e.g.:

> 1Kgs 8:60, כי יהוה הוא האלהים
> "That YHWH is God!"

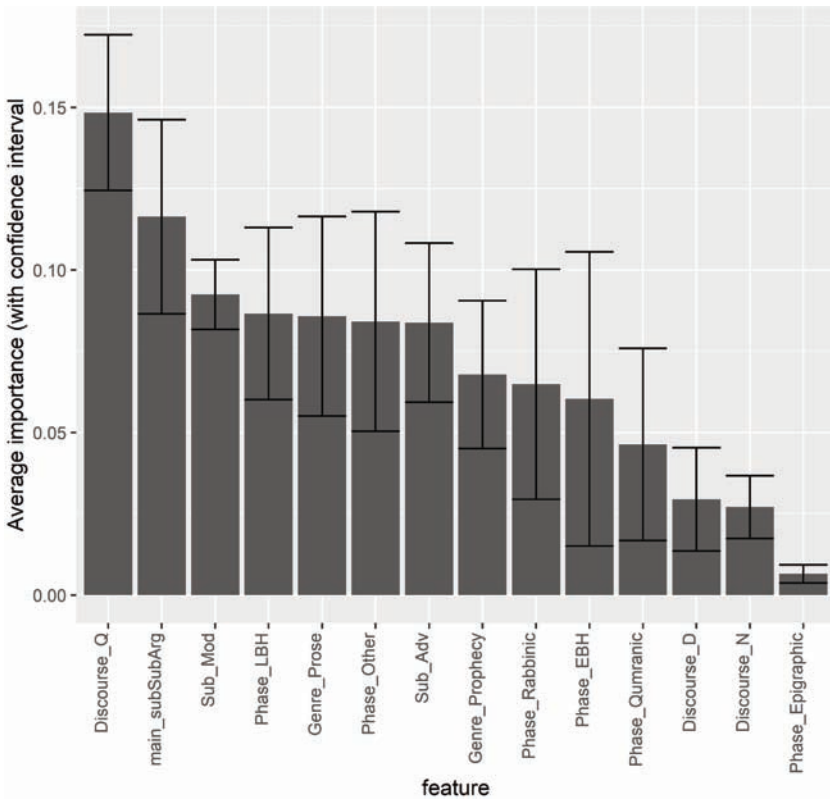

FIGURE 4.44 *Variable importance of the XGBoost model*

followed by a contrast:

> אין עוד
> "There is no other."[62]

62 Other cases in the Hebrew Bible are: Deut 4:35, לדעת כי יהוה הוא האלהים אין עוד מלבדו, "To know that YHWH is God; there is no other besides him.";

Deut 4:39, האלהים בשמים ממעל ועל־הארץ וידעת היום והשבת אל־לבבך כי יהוה הוא מתחת אין עוד, "So know today and take to heart that YHWH is God in heaven above and on the earth beneath, there is no other.";

Deut 7:9, וידעת כי־יהוה אלהיך הוא האלהים האל הנאמן שמר הברית והחסד לאהביו ולשמרי מצותו לאלף דור, "And you know that the LORD your God is God, the faithful God who maintains covenant loyalty with those who love him and keep his commandments, to a thousand generations.";

These two variables are the most important ones in the XGBoost model. After these, there is a whole range of lesser contributing variables with more or less equal importance. The confidence intervals indicate that it is difficult to say whether there is a real difference in importance between them. Among these are the various levels of the variables genre and language phase. Even though the levels of these variables contribute a bit to the predictive strength of the model, they do less so than the two most important ones.

Discussion of the results

The number of tripartite clauses in the Hebrew Bible is relatively low in comparison with the alternative bipartite construction. Muraoka already concluded that it would be more likely that if it functions as a copula, one would expect a higher frequency of them. This seems to be a strong argument. Although there are no concrete numbers available, it looks like the tripartite construction occurs much more often in Syriac than in BH (KP 331, Dyk and Van Keulen 2013: 409–411). Based on this, it is not only plausible that the pronoun functions as a copula in Syriac (even though Muraoka is also sceptical in the case of Syriac and other Semitic languages), but without the evidence from other languages, maybe one would not even hypothesize that the pronoun functions as a copula in BH, because there is another, more obvious explanation.

HJ supposes that resumptive and copular cases can both be found in the Hebrew Bible. This seems reasonable. If the pronoun can function as a copula, it is likely that it can still function in other clauses as a resumptive pronoun, because resumption is a widespread phenomenon in BH. However, the result is that only a small minority of the tripartite clauses function as a copular clause, which makes Muraoka's argument only stronger.

Deut 9:3 וידעת היום כי יהוה אלהיך הוא־העבר לפניך אש אכלה, "Know then today that the LORD your God is the one who crosses over before you as a devouring fire";

1 Kgs 20:31, הנה־נא שמענו כי מלכי בית ישראל כי־מלכי חסד הם, "Look, we have heard that the kings of the house of Israel are merciful kings"; 1 Kgs 8:60, למען דעת כל־עמי הארץ כי יהוה הוא האלהים אין עוד, "So that all the peoples of the earth know that the LORD is God; there is no other."; 2 Kgs 4:9, הנה־נא ידעתי כי איש אלהים קדוש הוא עבר עלינו תמיד, "Look, I know that this man who regularly passes our way is a holy man of God";

Isa 30:7, לכן קראתי לזאת רהב הם שבת, "Therefore I have called her, Rahab who sits still.";

Ps 100:3, דעו כי־יהוה הוא אלהים הוא־עשנו, "Know that the YHWH is God, he made us";

Qoh 1:17, ידעתי שגם־זה הוא רעיון רוח, "I know that this also is but a chasing after wind.";

Ezra 2:59, להגיד בית־אבותם וזרעם אם מישראל הם, "To tell whether their fathers and their seed are from Israel."

The most important variables in the XGBoost model clearly show the environments in which the tripartite verbless clauses occur most: these are quoted speech and clauses functioning as the argument of another clause, mostly as object. The most obvious explanation for these results is that the resumptive pronoun in these clauses puts emphasis on or gives prominence to the resumed subject.

It is true that the tripartite clause occurs in a variety of constructions, just like bipartite clauses and היה clauses, but the preferred environments of tripartite clauses differ from those of היה clauses. The environment of tripartite clauses generally is more similar to the environment of יש clauses: relatively often, they are arguments of another clause and they occur generally in quoted speech. This does not necessarily mean that the pronominal copula does not exist in BH. The most likely candidates are the cases in which there is incongruence between the pronoun and the resumed subject (e.g., Isa 37:16 אתה־הוא האלהים). This group is only a small subset of all the tripartite clauses in BH, and as we have seen, alternative explanations are possible, such as the interpretation of the pronoun functioning as a particle. The question is, of course, what is needed to give these cases the status of a distinct grammatical feature. Given that these cases are infrequent, and occurring generally in quoted speech, the most likely explanation is something other than that of a pronominal copula.

4.5. Conclusions

In this chapter, four ways of BH expressions of "to be" were analyzed. These four ways are the bipartite and tripartite verbless clause, and clauses with היה and יש. The most frequent by far of these is the bipartite clause, which functions as a kind of standard structure in this research. The analysis was done in three sections, and in each section verbless bipartite clauses were contrasted with one of the other three types. In sections 4.2, 4.3, and 4.4, the contrasting clauses were clauses containing היה, יש, or a resumptive pronoun respectively. Each of the three analyses required its own way of preparing and analyzing the data to make a good comparison possible.

Characteristic of the three datasets that were prepared for the analyses in this chapter is that they are imbalanced with respect to the output variable. In all the datasets, the ordinary bipartite verbless clause occurs more often than the syntactic alternative. In the case of the dataset with היה clauses, the less frequent value, clauses containing היה, occurs more than 1,000 times, but in the other datasets the less

frequent alternative occurs in a lower frequency, namely, 67 and 201 for, respectively, clauses containing יש and tripartite clauses. For this reason, it was decided that the analysis of the latter two datasets was done using a different approach than the היה clauses.

4.5.1. היה

The variation between היה clauses and bipartite verbless in section 4.2 clauses was analyzed using a Generalized Additive Mixed Model (GAMM). This is an extension of an ordinary logistic regression model. In logistic regression, the conditioning of a binary output is modeled. A mixed model takes into account that the observations within an individual book are not independent of each other. This is an important improvement, as neglecting the difference between fixed and random effects can lead to lower p-values than justified. The other two analyses, in sections 4.3 and 4.4, were done using Random Forest and Extreme Gradient Boosting. These are algorithms that are widely used in the field of predictive analytics for structured data.

Generally, the use of the verb היה is associated with the addition of TAM to a clause. This is confirmed in this research. The most direct evidence comes from clauses including time phrases. In these clauses, there is a significantly higher use of היה than in clauses without time phrase. The presence of phrases with certain functions has significant effect on the use of היה: clauses containing question and interjection phrases have a lower use of היה than clauses without these phrases; clauses with question and interjection phrases occur predominantly in quoted speech, but this decreased use of היה in these clauses is not visible in quoted speech in general, because there is no significant difference between the different levels of the variable discourse. היה is used less in poetry than in the other genres of prose and prophecy. Generally, poetry is more "timeless" than narrative, which coheres well with the idea that היה adds TAM to the clause.

The embedding of a clause in the surrounding text also influences the use of היה. Clauses containing conjunction phrases have an increased use of היה, and this is also the case for clauses with a mother which is not a verbless clause. Related to this is the decreased use of היה in subordinate clauses. In all three types of subordinate clauses, there is a strongly reduced use of היה as compared to main clauses. Also related to this is the increased use of היה if the mother of a clause contains a (finite) verb.

It is surprising that LBH has a significantly lower use of היה than EBH. This effect is mainly visible in the smaller LBH books (Esther, Daniel, Ezra, Nehemiah), in which

clauses with subject and predicate complement nearly always occur without the verb. The effect is not visible in other late literature (DSS and RH), so LBH seems to behave distinctively here.

In longer clauses, there is an increased use of היה. This result suggests that היה is not only used to add TAM to the clause, but also that it structures the clause. There is a certain collinearity of clause length and the presence of time and conjunction phrases. If a clause contains a time phrase it also means that it is longer than a clause without such a phrase. On the other hand, however, there are clauses with phrases with specific functions (question and interjection phrases) that have a reduced use of היה, this collinearity is not a problem in the present analysis.

The verb היה is used more in clauses with a PP than in clauses with a NP, AdjP or IPrP as predicate complement. The most likely explanation for this is that היה is used with the meaning "to happen" more often in the case of a PP predicate complement, which requires the use of the verb.

4.5.2. יש

There is a strong contrast in the use of יש between quoted speech and narrative text. Nearly all the cases of the particle can be found in quoted speech. This supports the idea that the particle adds emphasis to the clause, or part of it. This overrepresentation of יש in quoted speech is by far the most important categorical predictor for the use of the particle, as became clear in the XGBoost analysis. יש occurs relatively often in the selected Rabbinic texts. However, most of these clauses are clauses in quoted speech, and the RH subcorpus that is used here is too small to say that there is an increased use of the particle in RH, independent of its use in quoted speech.

4.5.3. *Tripartite clauses*

The most important variables in the XGBoost model with preference for tripartite verbless clauses are quoted speech (Q) and argument clauses. These results suggest that the resumptive pronoun emphasizes the resumed subject, because these environments are an obvious place for emphasis. This is strengthened by the number of tripartite verbless clauses in the Hebrew Bible. In total, there are fewer than 200 cases in the MT, while the alternative bipartite construction construction occurs thousands of times. Of course, there could be some cases in which the pronoun functions as a copula, as is suggested by HJ, but the general picture shows that the emphasis hypothesis is much more plausible than the copula hypothesis.

4.5.4. *Length of clauses and phrases*

In both the analysis of היה/verbless clauses and the clauses with and without יש, the length of a clause seems to influence the presence of the element היה or יש. In the case of היה, the result is statistically strongly significant and has a more or less linear pattern: in longer clauses there is a higher attestation of היה. The same pattern, although less clear, is visible for clauses with יש.

The length of the subject seems to play a role in the attestation of יש and the resumptive pronoun. Both seem to occur relatively more often with longer subjects,[63] and יש also with short subjects consisting of a single word. So, next to the semantic role of היה, יש, and the resumptive pronoun, they seem to have a role in structuring the clause. The role of particles in structuring a clause was suggested already by others, but it is the quantitative argument which strengthens this conjecture, by simply measuring how often a particle occurs in a clause or with a subject of a certain length, and comparing it with the alternative clauses without the particle.

4.5.5. *Diachronic variation*

Concerning the language phase, there are some interesting things to note. In LBH, היה occurs infrequently in the construction with a subject and predicate complement. LBH is actually the only level of the variable language phase which differs significantly from the base level EBH. Does this result imply a diachronic shift in the use of היה? In various studies, the increased use of היה with a participle is thought to be a late development, especially because it occurs often in RH. The decreased use of היה with subject and predicate complement is not seen in later language phases, so this decrease could be a phenomenon which occurs in LBH only. Of course, only a few post-biblical texts are taken into account in this research, so future research with more texts could shed new light on these results.

Similarly, going from EBH to LBH, there is a decrease in the use of the tripartite clause. This low frequency of the tripartite clause is also seen in QH, but not in RH, so there is not some kind of linear trend of this phenomenon visible in the history of Hebrew.

There does not seem to be an increased or decreased use of יש between EBH and LBH, so I cannot confirm the suggestion of Schülze, that its use increased during

63 In the case of the tripartite clause, this was not observed in this research, but by Driver (1892a: 268).

the history of BH. For further diachronic research about the use of היה, יש, and the resumptive pronoun, it is recommended to use more postbiblical texts than in this research. The availability of QH and RH texts for large scale research is still limited, but I expect that much more well-prepared texts will be available in the near future. Also, it is recommended to study interactions between variables. How do clause length and language phase interact? Likewise, studying redactional layers, literary subgenres, and parallel texts may shed light on diachronic developments.

CHAPTER 5

Verbal valence

5.1. Introduction

In verbal clauses, a verb is considered to be the core of the clause. The verb determines which other elements may accompany it. For instance, in English, a subject is part of the indicative clause, as in "I walk". Other verbs can be accompanied by more elements, for instance, the verb "to eat" can have a direct object, as in the clause "I eat bread". A verb that can take a direct object is called a transitive verb. Verbs of movement, like "to walk", that do not have a direct object, are called intransitive verbs.

Verbal valence or simply valency refers to the number of arguments a verb can take. The term valence was borrowed from chemistry by the French linguist Tesnière (1969). Dixon and Aikhenvald (2000: 2) distinguish between core and peripheral elements of a clause (examples are borrowed from Dixon and Aikhenvald 2000: 2):

> (On Monday morning), (in the garden), [John] danced (around the fountain).
> (On Monday morning), (in the garden), [a monkey] bit [John] (on the finger).
> (On Monday morning), (in the garden), [John] gave [Mary] [a book] (for her birthday).

Between square brackets, one finds core constituents of the clause required by the verb; these are also called complements of the verb. The peripheral elements are indicated by parentheses. These peripheral elements, which can be omitted without causing ungrammaticality, are called adjuncts. There are differences between languages as to what exactly should be understood as "required". For instance, in Latin, verbs are strictly transitive or intransitive (Dixon and Aikhenvald 2000: 4), but in many languages, words can be used in both transitive and intransitive clauses, e.g., in English the above mentioned transitive verb "to eat" can occur with and without a direct object.

Valence does not only deal with the transitivity of a verb. Dyk, Glanz, and Oosting (2014: 46, examples are borrowed from this publication) distinguish various valence

patterns. In the first place, there are impersonal or avalent verbs. In the expression "it rains", the subject is used to fulfill syntactical requirements without providing explicit meaning. Intransitive or monovalent verbs occur with a subject, which is required in languages like English. Transitive or divalent verbs have a subject and an object, like in the above-mentioned clause "I eat bread". Ditransitive or trivalent verbs have a subject, a direct object, and an indirect object, like the verb "to give" in the example "John gave Mary a book". There are even tetravalent verbs, for instance, in the English example "The fool bet him five quid on 'The Daily Arabian' to win". This clause has a subject ("the fool"), a direct object ("five quid"), an indirect object ("him") and a complement ("The Daily Arabian").

Examples of polyvalent verbs in BH are נתן and שים. They can occur without a direct object, but also with one or two objects in various constructions. In these different constructions, the verbs need to be rendered differently in translation. The meaning of נתן and שים can be synonymous in so-called double object constructions. This synonymity is observed in grammars and lexicons, but no research has been done on how the choice for one of the alternative lexemes is conditioned. Therefore, the research question here is:

> Which variables influence the choice of the lexemes נתן and שים in constructions with similar syntactic and semantic content, when the verb occurs with a double object construction?

In this chapter, first an impression will be given of the research on valence research in Hebrew (section 5.2), then the polyvalent verbs נתן and שים will be introduced in section 5.3. In section 5.4, the double object patterns of these verbs will be studied from a quantitative perspective.

5.2. Valence research in Hebrew

Various studies related to verbal valence in BH have been published, in which it is the main focus of research (Dyk, Glanz, and Oosting 2014; Glanz, Oosting, and Dyk 2014; Oosting and Dyk 2017; Winther-Nielsen 2017; and especially Malessa 2006). For Oosting (2011: 26–27), the study of the valence of the verb is an important starting point in the syntactic analysis. In relation to linguistic variation and diachrony in BH, the focus of valence research has been mainly on three aspects:

5.2.1. *Variation between verb forms*

In his work on the language of Chronicles, Kropat (1909: 14–15) focuses on places where Chronicles deviates from its supposed *Vorlage*. He gives a number of examples where Chronicles uses a transitive form, where the parallel passage in Samuel or Kings uses an intransitive alternative:

1 Sam 31:12 ויבאו יבשׁה
And they came to Jabesh.

1 Chr 10:12 ויביאום יבישׁה
And they brought them to Jabesh.[1]

After these examples Kropat gives a number of examples, where active verb forms occur in Chronicles where the EBH parallel uses passive forms:[2]

2 Sam 5:13 ויולדו עוד לדוד בנים ובנות
And more sons and daughters were born to David.

1 Chr 14:3 ויולד דויד עוד בנים ובנות
And David begat more sons and daughters.

Kropat also gives the only case where Chronicles uses a passive form where the parallel in Samuel/Kings uses an active form:

2 Sam 5:17 כי־משחו את־דוד
That they anointed David.

2 Chr 14:8 כי־נמשׁח דויד
That David was anointed.

1 The other parallels given by Kropat are 2 Sam 6:9 // 1 Chr 13:12, 2 Kgs 22:9 // 2 Chr 34:16, 2 Kgs 23:34 // 2 Chr 36:4, 2 Sam 7:15 // 1 Chr 17:13, 1 Kgs 10:29 // 2 Chr 1:17.

2 The other parallels given by Kropat are 2 Sam 7:16 // 2 Chr 17:14, 1 Kgs 11:43 // 2 Chr 9:31, 2 Kgs 14:20, 15:38, etc. // 2 Chr 25:28, 27:9, etc., 2 Kgs 11:2 // 2 Chr 22:11, 2 Kgs 11:15, 16 // 2 Chr 23:14,15.

Kropat suggests that Chronicles generally has a preference for active and transitive verb forms, relative to its *Vorlage* Samuel and Kings. However, he does not say explicitly that this is a general tendency of later language.

5.2.2. *The use of the direct object (with* את *or alternative constructions)*

In BH, the most frequently used particle to introduce the direct object is את, but its use is irregular. Even in clauses with a more or less identical meaning there is variation in its use, even within the same book (Malessa 2006: 27 gives more examples):

> Exod 7:26 שַׁלַּח אֶת־עַמִּי
> Exod 8:16 שַׁלַּח עַמִּי
> Let my people go.

However, there are some factors that seem to influence the use of את. It is used more or less exclusively in definite phrases (Malessa 2006: 33), though this does not explain the variation of the use of את within the set of definite phrases. According to Malessa the highest use of את can be found in phrases that represent a human being. He proposes a semantic hierarchy in which there is a decreasing use of את when the phrase introduced functions as direct object (Malessa 2006: 33).

> [human being] ▸ [living creature] ▸ [concrete] ▸ [abstract]

This hierarchy can also be found when the object is a demonstrative pronoun. The object is clearly more often introduced by את if the pronoun refers to a human being rather than to abstract nouns (Malessa 2006: 34). Abstract phrases like הרע ("the evil") and הטוב ("the good") as object of the verb עשה ("to make") are rarely marked with את (Malessa 2006: 39; Wilson 1889).

Concerning variation between EBH and LBH, Malessa observed that in the parallel passages in Samuel, Kings, and Chronicles, there does not seem to be a consistent pattern. Sometimes Chronicles lacks את where it is used in Samuel/Kings, and sometimes vice versa (Malessa 2006: 60–61).

5.2.3. *Verbs of movement and their locatives*

In the literature on BH, the locative phrase functioning as complement of a verb of movement has been described extensively. It is a complex topic, in which various

patterns of variation play a role. The movement indicated by the verb can be diverse, it can be a movement in space or time, but it can also be a figurative movement. The locative can indicate the origin, the traversed path, or the goal of the movement, and the goal can be a location or a person or both (Oosting and Dyk 2017: 2). There is a variety of movement verbs, and in the literature, generally, the most common are described. Oosting and Dyk (2017: 2) restrict themselves to verbs occurring 150 times or more in the MT. These are בוא ("to come"), הלך ("to go"), יצא ("to go out"), ירד ("to go down"), נוס ("to flee"), נפל ("to fall"), סור ("to turn aside"), עבר ("to pass"), עלה ("to go up"), and שוב ("to return"). This is a semantically diverse group of verbs, and also their complements are diverse. The verbs occur mostly in the Qal, but also in the Hiphil and Hophal.

The complement can be introduced by a preposition. In his comprehensive study, Austel (1968, 18) restricts himself to complements starting with אל, ל, ב, על, and to complements not initiated by a preposition, sometimes called accusative. However, a wider range of combinations of prepositions is used, such as עד, מן, just like the so-called ה-locale, which occurs more than 1,000 times in the MT. Further, one can distinguish between different referents in the complement. The complement can refer to a (geographic) location or a person, and a clause can also contain a combination of these elements.[3]

5.3. Polyvalent verbs in BH, נתן and שים

5.3.1. *Introduction*

In BH, many verbs occur in a variety of syntactic patterns. Most verbs that are transitive in the Qal occur with one object, but if the verb is made causative it can govern two objects. In that case, the subject of the Qal becomes an object (GKC §117cc a; JM §125u). Some examples:

> Exod 33:18 הראני נא את־כבדך
> Show (Hiphil) me, then, your glory.

3 For more comprehensive explanantions with examples, I refer to Austel (1968); Hornkohl (2014: sections 7.3 and 7.4), Rezetko and Young (2014: section 9.4), and Oosting and Dyk (2017).

Deut 8:3 ויאכלך את־המן אשׁר לא־ידעת
He fed (Hiphil) you with manna, which you did not know.

There are Qal verbs that can take a double object (JM §125ua.):

Lev 24:23 וירגמו אתו אבן
And they stoned him [with] a stone.

Ezek 13:10 והנם טחים אתו תפל
And behold, they plaster it [with] mortar.

These double object verbs are often verbs of asking or desiring (WO §10.2.3; GKC §117 gg; e.g., Deut 14:26 or Ps 137:8), or answering (ענה in Mic 6:6; GKC §117 gg), verbs of clothing (GKC §117 ee; לבשׁ in 1Sam 7:38, or פשׁט in Gen 37:23), making something out of something (GKC §117 hh; e.g., Gen 2:7), or making something into something (GKC §117 ii; e.g., Gen 27:9).

A double object construction can be seen as a verb governing a so-called small clause (Haegeman 1991: 160–161; Dyk, Glanz, and Oosting 2014: 54) if the meaning of the clause is "turn something into something else", according to Glanz, Oosting, and Dyk (2015: 47):

Judg 17:4 ויעשׂהו פסל
And he made it [into] a statue.

1Sam 8:1 וישׂם את־בניו שׁפטים לישׂראל
And he made his sons judges over Israel.

The two objects together form a verbless clause.

The second object can be introduced by ל. In such a construction, the phase introduced by ל should not function as a manner of action, or as a location of action, and also it does not refer to the one affected by the action (Dyk, Glanz, and Oosting 2014: 54–55, referring to Haegeman 1991: 160–161). An example of such a double object construction, in which the second object is introduced by ל is Exod 32:10, ואעשׂה אותך לגוי גדול, "I will make you into a great nation". The two objects must refer to the same participant in order to qualify as a double object construction.

For JM, the basis of this kind of double object clause is a היה clause, in which היה has the meaning "to become" (JM §125 w). According to JM (§125 v) the following type of clause is derived from a nominal clause:

Gen 2:7 וייצר יהוה אלהים את־האדם עפר מן־האדמה
And the Lord God formed man [from] dust from the earth.

The second object of the verbal clause is עפר, the material of which האדם was made. The subject of the nominal clause becomes the object in the verbal clause, and in the verbal clause the second object does not take the object marker את.

5.3.2. *Introduction to the valence of* נתן *and* שים

In Biblical Hebrew, a number of double object verbs occur in similar contexts, for example, the verbs נתן and שים. Glanz, Oosting, and Dyk (2015: 36–37) give the following examples that are often translated similarly:

Jer 9:10 ואת־ערי יהודה אתן שממה מבלי יושב
I will make the towns of Judah a desolation, without inhabitant.

Jer 10:22 לשום את־ערי יהודה שממה מעון תנים
To make the cities of Judah a desolation, a lair of jackals.

They refer to BDB, according to which the meaning of נתן "nearly equals שים" (BDB, נתן Qal, section 2). However, it is not always necessary to translate נתן and שים identically, because the words occur in different contexts. שים is often found with contrasting elements (e.g., Isa 5:20; 41:18; 42:16; 50:2; Job 17:12; Ps 107:33; 35), which makes it possible to translate שים with "to turn into" (Glanz, Oosting, and Dyk 2015: 50).

Another example of near synonyms are שים and קרא. Both verbs are used in the context of giving names, but in the case of שים someone is often renamed, whereas the verb קרא is generally used if someone gets a name for the first time (Glanz, Oosting, and Dyk 2015: 50). Some examples are:

Gen 3:20 ויקרא האדם שם אשתו חוה
The man named his wife Eve.

Gen 4:25 ותקרא את־שמו שת
And she named him Seth.

2 Kgs 17:34 אשר־שם שמו ישראל
Whom he named Israel.

Dan 1:7 וישׂם להם שׂר הסריסים שׁמות
The palace master gave them [other] names.

Even though the verbs נתן and שים may be semantically similar in some constructions, there may also be substantial differences. נתן and שים are both multivalent verbs.[4]

If נתן occurs without complements, generally a direct object is implied.[5] In these cases, the verb has the same meaning with one complement, if this is a direct object.[6] There are also other clauses with נתן plus a single complement, such as an indirect object. In the latter case, a direct object is assumed to be implied, if it has been mentioned in the context already,[7] which is also the case if the verb is accompanied by a locative, although the meaning of the verb then differs.[8]

A frequent construction with two complements is a clause in which the verb נתן is accompanied by a direct and indirect object;[9] it can also occur with a direct object and a locative.[10] The patterns of נתן with two objects are discussed in section 5.3.3.[11]

An overview of the different valence patterns and their meanings is shown in figure 5.1 (see next page), which is a flowchart in which the valence of the verb נתן is linked to the different meanings that the verb can have (Dyk 2016, figure used with permission).

The verb שים has a similar spectrum of valence patterns. There are three main groups of valence patterns of the verb שים. These are clauses with the verb שים in which:

- there is no direct object in the clause.
- there is one direct object and a locative.
- there are two objects (double object clauses).

---

4 For an overview, see https://github.com/ETCBC/valence/wiki.

5 E.g., Gen 30:28, ואתנה, "And I will give"; Job 1:21, יהוה נתן, "YHWH has given".

6 E.g., Ps 18:14, ועליון יתן קלו ברד וגחלי־אשׁ, "And the highest gave his voice, hail and coals of fire"; 2 Chr 34:9, ויתנו את־הכסף, "And they gave the money".

7 E.g., Gen 20:14, ויתן לאברהם, "And he gave (it) to Abraham".

8 E.g., Exod 12:7, ונתנו על־שׁתי המזוזת ועל־המשׁקוף על הבתים, "And they put (it) on the two side posts and on the upper door posts of the houses"; 2 Kgs 5:23, ויתן אל־שׁני נעריו, "And he put (them) on two of his boys".

9 E.g., Gen 34:9, בנתיכם תתנו־לנו, "Give your daughters to us"; Judg 17:10, ואנכי אתן־לך עשׂרת כסף לימים, "And I will give you ten pieces of silver by the year".

10 E.g., Dan 1:2, ויתן אדני בידו את־יהויקים מלך־יהודה ומקצת כלי בית־האלהים, "And the lord put in his hand Jehoiakim, king of Judah and part of the vessels of the house of God"; Jer 31:33, נתתי את־תורתי בקרבם, "I will put my law in their midst".

11 The variation in the meaning in relation to valence of the verb נתן is described by Winther-Nielsen (2017), from a different perspective, namely, that of Role and Reference Grammar.

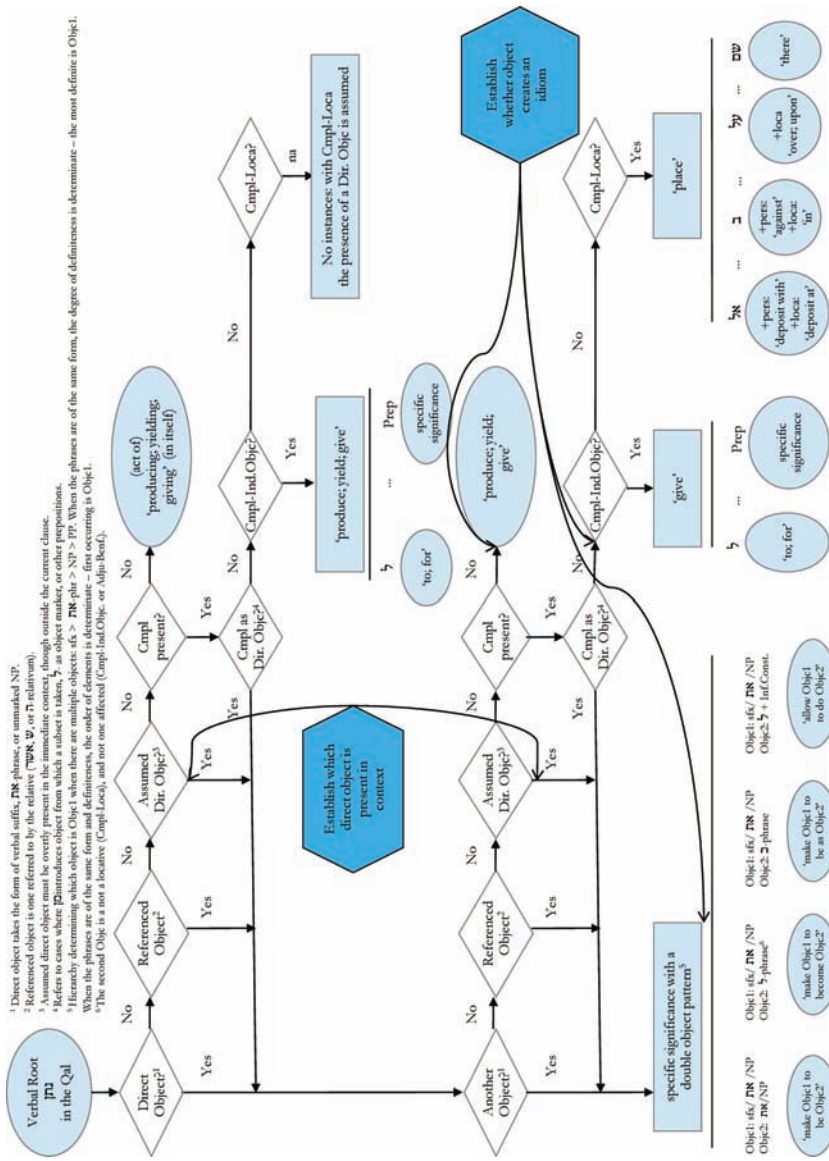

FIGURE 5.1 *Flow chart of the different valence patterns and meanings of the verb* נתן

The verb and the semantic variation between the various valence patterns are described concisely in the flowchart published in Dyk (2016, figure copied with permission), see figure 5.2 (see next page). The verb שים can be found without complements (although a direct object can be implied, if that can be derived from the context),[12] but also with one complement, which can be a locative[13] or a direct object.[14] If it occurs with two complements, these can be a direct object and a benefactive adjunct,[15] or a direct object and a locative.[16]
The double object patterns of the verb שים are discussed in section 5.3.3.

5.3.3. *Exploration of double-object patterns of* נתן *and* שים

Both נתן and שים occur with a variety of double object constructions. נתן can occur with a direct object and an infinitive object. In this construction, the first object is the subject of the action of the verb in the infinitive object clause:

> Exod 3:19 ואני ידעתי כי לא־יתן אתכם מלך מצרים להלך
> And I know, that the king of Egypt will not let you go.

> Exod 12:23 ולא יתן המשחית לבא אל־בתיכם
> And He will not allow the destroyer to enter your houses.

The verb שים does not occur with this construction.

Both נתן and שים can occur with two direct objects. The meaning of both constructions in most cases is "to make object_1 to be object_2".

> Deut 1:15 ואתן אתם ראשים עליכם
> I made them heads over you.

---

12 E.g., Isa 41:20, וישׂימו, "And they may consider".

13 E.g., Exod 2:3, ותשׂם בסוף על־שׂפת היאר, "And she placed (it) among the reeds on the bank of the river".

14 E.g., 1Sam 21:7, לשׂום לחם חם ביום, "To place hot bread on the day".

15 E.g., Gen 4:15, וישׂם יהוה לקין אות, "And YHWH appointed a mark for Cain"; 1Chr 17:9, ושׂמתי מקום לעמי ישׂראל, "I will appoint a place for my people Israel".

16 E.g., Gen 2:8, וישׂם שׂם את־האדם, "And there he put the man"; Ezek 4:2, ושׂים־עליה כרים סביב, "And put rams against it all around."

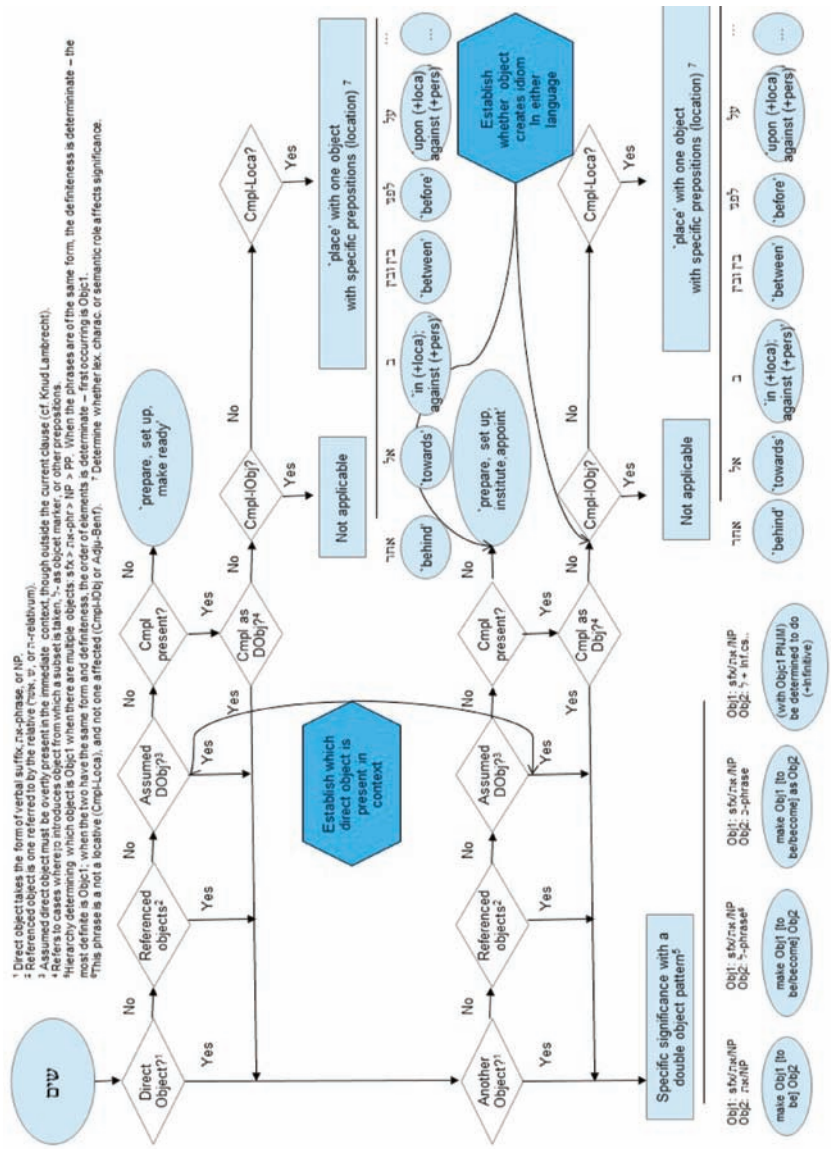

FIGURE 5.2 *Flowchart of the valence patterns and meanings of the verb* שים

Ps 2:8 ואתנה גוים נחלתך
And I will make the peoples your inheritance.

Gen 47:6 ושׂמתם שׂרי מקנה על־אשׁר־לי
And make them rulers of the cattle, over that which is mine.

Isa 50:2 אשׂים נהרות מדבר
I make the rivers a wilderness.

Closely related to the former pattern is the pattern with verb + direct object + ל-object, which also occurs with both נתן and שׂים.

Gen 17:6 ונתתיך לגוים
And I will make you to peoples.

1Kgs 2:17 ויתן־לי את־אבישׁג השׁונמית לאשׁה
And he will make Abishag the Shunnamite a wife for me.[17]

Zech 7:14 וישׂימו ארץ־חמדה לשׁמה
And they made a pleasant land a desolation.

Gen 21:13 וגם את־בן־האמה לגוי אשׂימנו
As for the son of the slave woman, I will make him a nation.

נתן and שׂים can also occur with a direct object and a כ-object. This construction means "to make direct object like כ-object".

Lev 26:19 ונתתי את־שׁמיכם כברזל
And I will make your sky like iron.

17 There is an alternative way of interpreting this clause and the construction of נתן x לאשה. In this interpretation, the wife is seen as the possession of the man, and the clause in 1Kgs 2:17 would be translated as "He gives to me Abishag the Shunnamite, as wife". Then, לאשה functions as adjunct, and the clause has only a single object. I have chosen to count these constructions as double object clauses.

1 Kgs 16:3 ונתתי את־ביתך כבית ירבעם בן־נבט
And I will make your house like the house of Jeroboam son of Nebat.

Gen 13:16 ושׂמתי את־זרעך כעפר הארץ
I will make your offspring like the dust of the earth.

Job 41:23 ים ישׂים כמרקחה
It makes the sea like a pot of ointment.

Parallel passages

There are various attestations of double object constructions of נתן and שים occurring in parallel passages in the MT. Most of these parallel attestations have identical constructions, but it is interesting to compare cases in which they are different. There is a case, in which Chronicles uses נתן, and the parallel in Samuel uses כון.

1 Chr 17:22 ותתן את־עמך ישׂראל לך לעם עד־עולם
And you made (נתן) your people Israel to be a people for yourself forever.

2 Sam 7:24 ותכונן לך את־עמך ישׂראל לך לעם עד־עולם
And you established (כון) your people Israel for yourself to be a people for yourself forever.[18]

Also, there is a case in which Chronicles uses נתן and the parallel clause in Kings uses שים.

2 Chr 9:8 ויתנך עליהם למלך לעשׂות משׁפט וצדקה
He has made (נתן) you king over them, that you may execute justice and righteousness.

1 Kgs 10:9 וישׂימך למלך לעשׂות משׁפט וצדקה
He has made (שים) you king to execute justice and righteousness.

18 The different choice of verbs is also reflected in the LXX text: 1 Chr 17:22, καὶ ἔδωκας τὸν λαόν σου Ἰσραὴλ σεαυτῷ λαὸν ἕως αἰῶνος; 2 Sam 7:24, καὶ ἡτοίμασας σεαυτῷ τὸν λαόν σου Ἰσραὴλ εἰς λαὸν ἕως αἰῶνος.

Both examples in Chronicles use the verb נתן with direct object and a ל-object, and in both cases the parallel uses a different verb (כון and שים), without having a clearly different meaning. Another interesting parallel can be found in the same verses as the latter example:

2 Chr 9:8 לתתך על־כסאו למלך ליהוה אלהיך
… to make you king on the throne for the Lord your God.

1 Kgs 10:9 לתתך על־כסא ישׂראל
… to put you on the throne of Israel.

Here Chronicles has a double object construction, where Kings has a single object. In 2 Chr 9:8, there are two double object constructions with the verb נתן, whereas Kings has one double object construction with שים, and a clause with נתן with a single object.[19] The clause לתתך על־כסאו למלך ליהוה אלהיך in 2 Chr 9:8 is more complicated than most double object clauses with נתן, because the locative על־כסאו stands between the two objects. An alternative interpretation of the clause is that it has the same meaning as the parallel in 1 Kgs 10:9; then, למלך would be an adjunct. I follow the interpretation in which it is a double object construction.

There is also a case in which Chronicles uses a double object construction with נתן, whereas the parallel in Kings uses the verb היה with an explicit subject.

2 Chr 7:20 ואתננו למשׁל ולשׁנינה בכל־העמים
And I will make it a proverb and a taunt among all the people.

1 Kgs 9:7 והיה ישׂראל למשׁל ולשׁנינה בכל־העמים
And Israel will be a proverb and a taunt among all the people.[20]

19 This variation in choice of verb and valence pattern is reflected equally in the LXX text: 2 Chr 9:8, ἔστω Κύριος ὁ Θεός σου εὐλογημένος, ὃς ἠθέλησεν ἐν σοὶ τοῦ δοῦναί σε ἐπὶ θρόνον αὐτοῦ εἰς βασιλέα Κυρίῳ Θεῷ σου· ἐν τῷ ἀγαπῆσαι Κύριον τὸν Θεόν σου τὸν Ἰσραὴλ τοῦ στῆσαι αὐτὸν εἰς αἰῶνα καὶ ἔδωκέ σε ἐπ᾽ αὐτοὺς εἰς βασιλέα τοῦ ποιῆσαι κρίμα καὶ δικαιοσύνην; 1 Kgs 10:9, γένοιτο Κύριος ὁ Θεός σου εὐλογημένος, ὃς ἠθέλησεν ἐν σοὶ δοῦναί σε ἐπὶ θρόνου Ἰσραήλ· διὰ τὸ ἀγαπᾶν Κύριον τὸν Ἰσραὴλ στῆσαι εἰς τὸν αἰῶνα καὶ ἔθετό σε βασιλέα ἐπ᾽ αὐτοὺς τοῦ ποιεῖν κρίμα ἐν δικαιοσύνῃ καὶ ἐν κρίμασιν αὐτῶν.

20 Also, in this case, the variation in the Hebrew is reflected in the LXX text: 2 Chr 7:20, καὶ δώσω αὐτὸν εἰς παραβολὴν καὶ εἰς διήγημα ἐν πᾶσι τοῖς ἔθνεσι; 1 Kgs 9:7, καὶ ἔσται Ἰσραὴλ εἰς ἀφανισμὸν καὶ εἰς λάλημα εἰς πάντας τοὺς λαούς.

Other parallels are 2 Chr 6:27 // 1 Kgs 8:36,[21] 2 Chr 7:20 // 1 Kgs 9:22,[22] 2 Chr 25:18 // 2 Kgs 14:9,[23] but in these cases נתן occurs in with identical constructions.[24]

The second object with and without ל and some idiomatic expressions.

Not only are the verbs נתן and שים semantically closely related in this kind of constructions, but the double object constructions with and without ל also seem to have no difference in meaning. For instance, in the following examples there is no clear semantic difference.

direct object + direct object
Deut 1:15 ואתן אתם ראשׁים עליכם
And I made them leaders over you.

direct object + ל-object
Deut 28:13 ונתנך יהוה לראשׁ
YHWH will make you a leader.

In books with many cases of double object constructions with the verbs under consideration, both second objects with and without ל seem to be used without a clear difference in meaning. That does not mean that the constructions with and without ל are always interchangeable. For instance, שים can be used when a name is given, but always without ל.

Judg 8:31 וישׂם את־שׁמו אבימלך
And he made his name Abimelech.

It must be said that it is difficult to say that only one alternative can occur if the dataset is relatively small. Therefore, it is important not to base strong conclusions on the observation that only one alternative can be found in the corpus.

Although in many cases the verbs נתן and שים seem to be interchangeable in double object constructions, there is a number of idiomatic expressions that occur

21 אשׁר־נתתה לעמך לנחלה, "Which you have made a heritage for your people".
22 ומבני ישׂראל לא־נתן שׁלמה עבד, "And Salomo did not make the people of Israel slave(s)".
23 תנה־את־בתך לבני לאשׁה, "Make your daughter for my son to be wife".
24 There is also a parallel with the book of Psalms: 1 Chr 16:18//Ps 105:11.

exclusively with one of the alternative verbs. The expression "to give someone as wife" occurs nearly exclusively with the verb נתן, for example, Josh 15:17, ויתן־לו את־עכסה בתו לאשה, "and he made for him his daughter Achsah as wife".[25] Another interesting characteristic of this expression is that the second object, the phrase with the wife, אשה, occurs nearly exclusively with ל.[26] The expression "to make someone's name x" does not occur with נתן, but varies between שים and קרא.[27]

5.3.4. *The data*

The data used in this research come from the valence dataset, produced by Janet Dyk and Dirk Roorda. This dataset is available in Text-Fabric format.[28] It is based on enrichments of the ETCBC database proposed by Dyk on the basis of research done in the "Data and Tradition" research project[29] (Dyk, Glanz, and Oosting 2014; Glanz, Oosting, and Dyk 2015; Oosting and Dyk 2017). The data from the GitHub repository have been cleaned manually.[30]

5.4. Quantitative analysis of נתן and שים with double object constructions

In this section, double object clauses with the verbs נתן and שים are analyzed in an explorative, quantitative way. First, the cases with a direct object followed by a second object with or without ל are discussed together, after that the cases in which the second object is introduced by כ will be analyzed.

---

25 Other occurrences are Gen 16:3, 29:28, 30:4, 30:9, 34:8 (Here the verb is nifal, so the subject is the logical object), 34:12, 38:14 (nifal), 41:45, Deut 22:16, Josh 15:16 and 17, Judg 1:12, 1:13, 21:1, 21:7, 1 Sam 18:17, 18:19 (nifal), 18:27, 1 Kgs 2:17, 2:21, 2 Kgs 14:9, and 1 Chr 2:35, 25:18.

26 An exception is 1 Kgs 11:19, ויתן־לו אשה את־אחות אשתו אחות תחפניס הגבירה, "And he made for him his sister-in-law for a wife, the sister of Tahpenes the mistress".

27 Other examples with שים can be found in Judg 8:31, 2 Kgs 17:34, Neh 9:7, in Aramaic in Dan 5:12.

28 The data can be downloaded from https://github.com/ETCBC/valence. Documentation on the datasets with a tutorial notebook can also be found there.

29 2010–2015, Investment grant from the Netherlands Organisation for Scientific Research (NWO) for the project *Bridging Data and Tradition. The Hebrew Bible as a Linguistic Corpus and as a Literary Composition* (application of E. Talstra and W.T. van Peursen, Vrije Universiteit Amsterdam and Leiden University).

30 The cleaned datasets can be found here: https://github.com/MartijnNaaijer/phdthesis/tree/master/Ch5_Verbal_valence.

As was already observed, נתן and שים have more or less the same meaning when the verb occurs in a clause with a double object construction. If that is true, the question arises why an author chooses one of the two verbs. While the why-question is always difficult to answer, especially in the case of ancient data, it is possible to describe the distribution of the verbs in double object constructions.

5.4.1. *Double object constructions of* נתן *and* שים *and the main variables of the Syntactic Variation project*

In section 5.3, some distinct cases of double object constructions with נתן and שים were investigated, but what is the situation in the whole of the MT? Do double object constructions with נתן and שים occur everywhere or only in specific books? Is there a preference for one of the two verbs?

A first impression is obtained by counting the total number of double object constructions of each of the verbs נתן and שים. Of all the occurences of the verb שים in the MT, nearly 21% are with a double object construction, whereas in the case of the verb נתן this is 10%.[31] How are these constructions distributed with respect to the main variables of this research?

Figures 5.3 and 5.4 (see next page) show the distribution of double object constructions versus other constructions of the verbs נתן and שים in main and subordinate clauses.

The mosaic plots clearly show the higher fraction of double object constructions of שים, relative to נתן. For both verbs, there are no double object constructions in which the clause functions as the argument in another clause. It might be the case that double object constructions are too complex to function as an argument in another clause.

In the case of the verb שים, double object constructions occur relatively more often in subordinate clauses than in main clauses. For נתן, this is the other way around, although the differences between main clauses and the two types of subordinate clauses are small.

Figures 5.5 and 5.6 show the distribution of double object constructions versus other constructions of the verbs נתן and שים in their discourse environment.

---

31 שים: 120 double object constructions to a total of 581 attestations of the verb, נתן: 205 double object constructions to a total of 2001 attestations of the verb in the MT. I do not deal with texts outside of the MT in this chapter. In her PhD thesis, Femke Siebesma discusses valence patterns of נתן, שים and other verbs in the DSS.

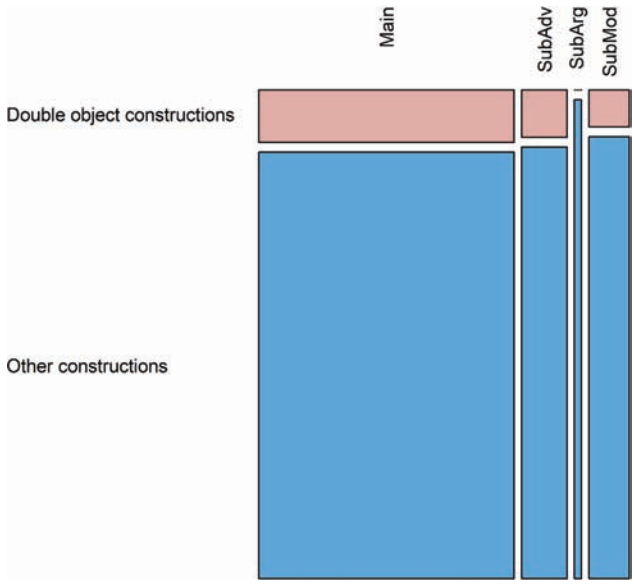

FIGURE 5.3 *Main and subordinate clauses with double object constructions versus other constructions of* נתן

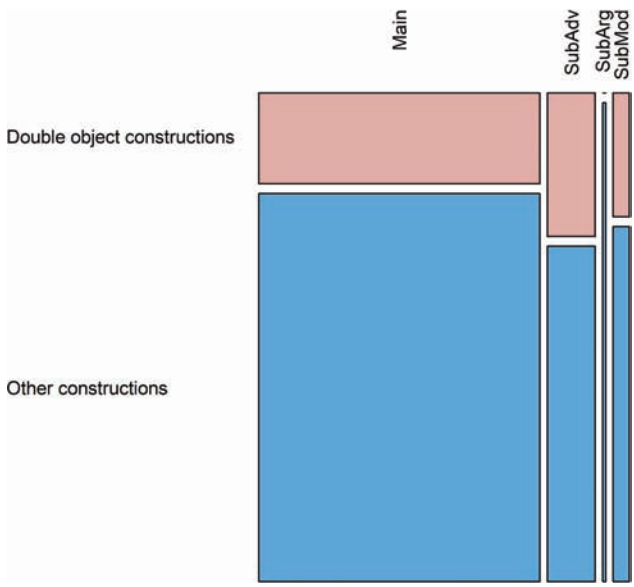

FIGURE 5.4 *Main and subordinate clauses with double object constructions versus other constructions of* שים

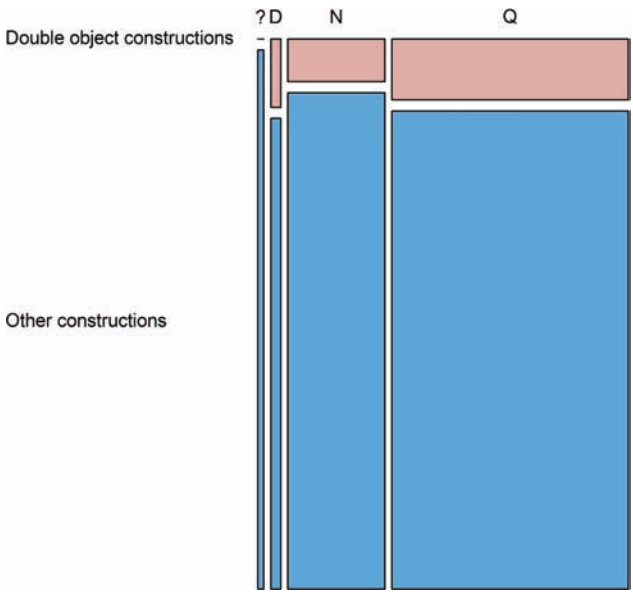

FIGURE 5.5 *Discourse type of double object constructions versus other constructions of* נתן

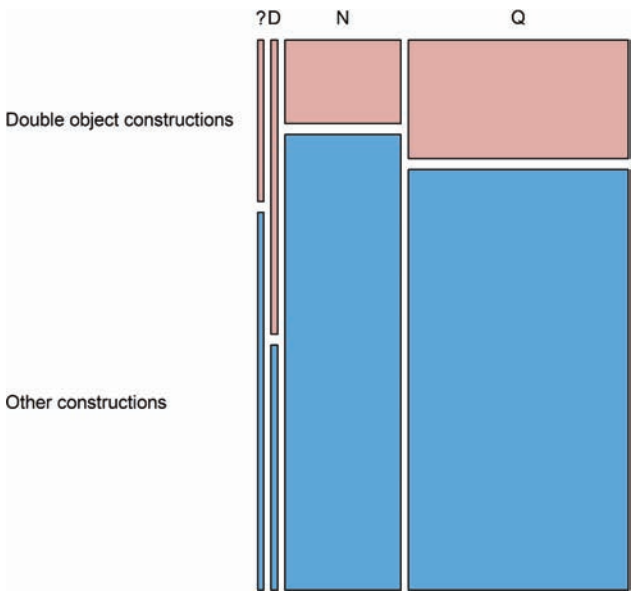

FIGURE 5.6 *Discourse type of double object constructions versus other constructions of* שים

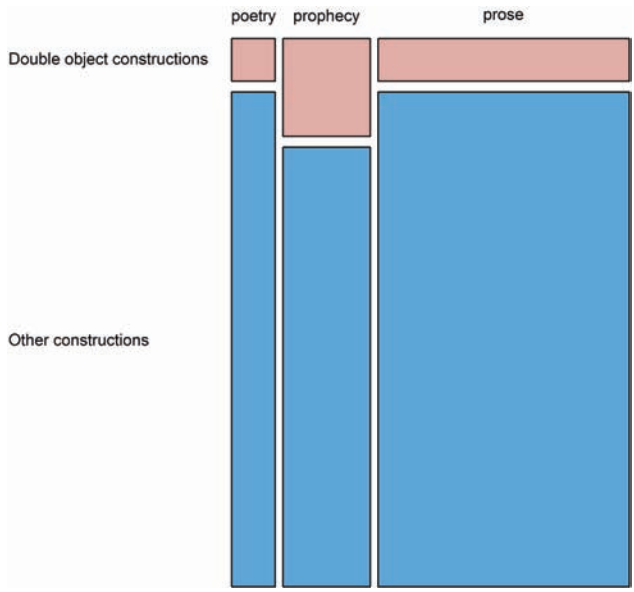

FIGURE 5.7 *Genre of double object constructions versus other constructions of* נתן

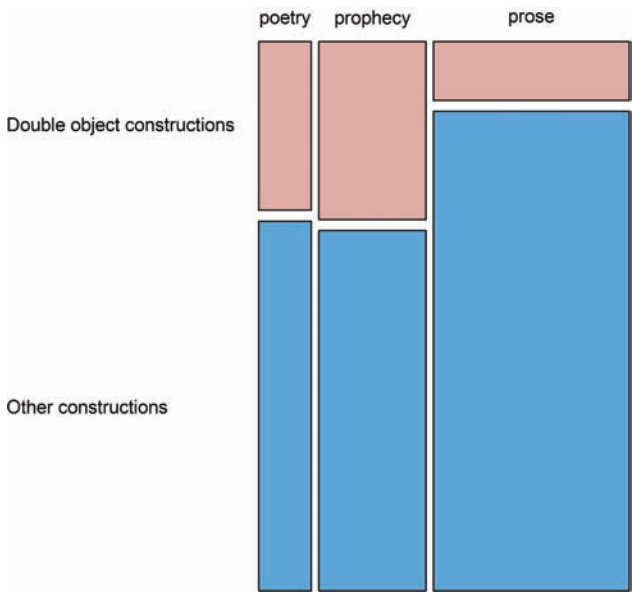

FIGURE 5.8 *Genre of double object constructions versus other constructions of* שים

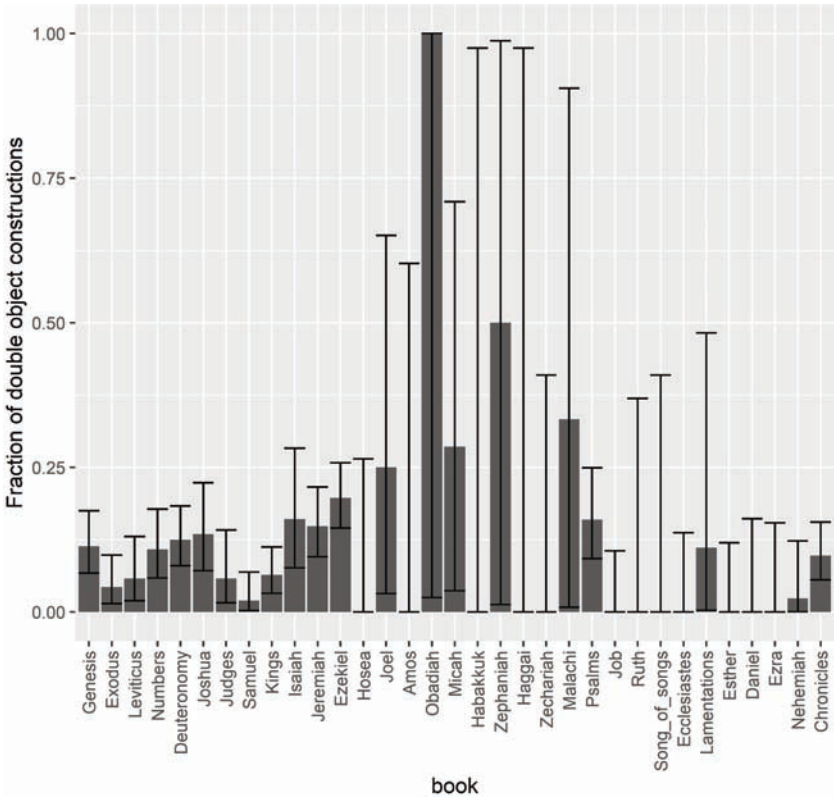

FIGURE 5.9 *Fraction of double object constructions with the verb* נתן

The levels ? and D are rare and less relevant, so they will be left out of consideration. The situation in Q and N is similar for both verbs: double object constructions occur more often in Q than in N clauses. This might be related to distribution of double object constructions in different genres, which is shown in figures 5.7 and 5.8 (see page 136). A comparison of figures 5.5 and 5.6 shows that double object constructions of נתן and שים are similarly distributed among Q and N.

For both verbs, the most frequent attestation of double object verbs can be found in the genre of prophecy, but for the verb שים, it occurs nearly as frequently in poetry. This does not mean that double object constructions are more or less absent in prose books, but there is a surprisingly high concentration in the other genres. Overall, double object constructions with נתן and שים are distributed similarly throughout the three genres under consideration, except that those with שים occur more in poetry than נתן.

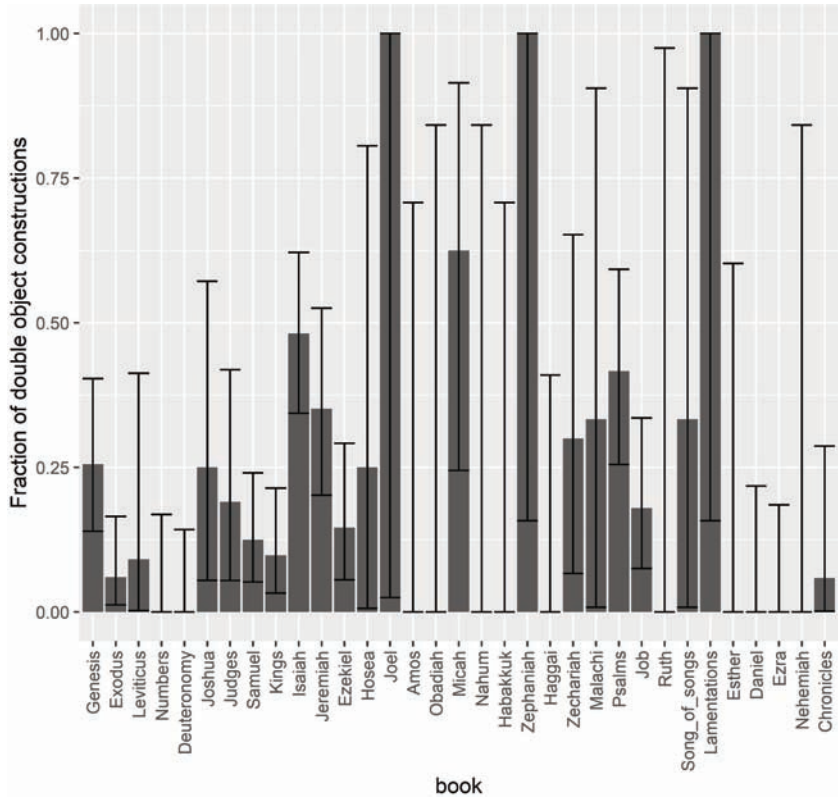

FIGURE 5.10 *Fraction of double object constructions with the verb* שים

What happens if we look at each book separately? Figure 5.9 (see page 137) and 5.10 show the fraction of the occurences of נתן and שים in each biblical book.[32] Here, the situation in EBH and LBH will be considered. The figures show double object constructions (with or without ל in the second object) as a fraction of the total amount of occurrences of נתן or שים, to get an impression whether certain books have a preference for using double object constructions relative to other books. As usual with this kind of data, there is substantial variation between individual books in the corpus. Most error bars in the figure are relatively wide, indicating that the height of the bar does not give much certainty about the preference for the double object construction in a book, especially in the shorter books of the MT,

32 The error bars are based on a binomial test.

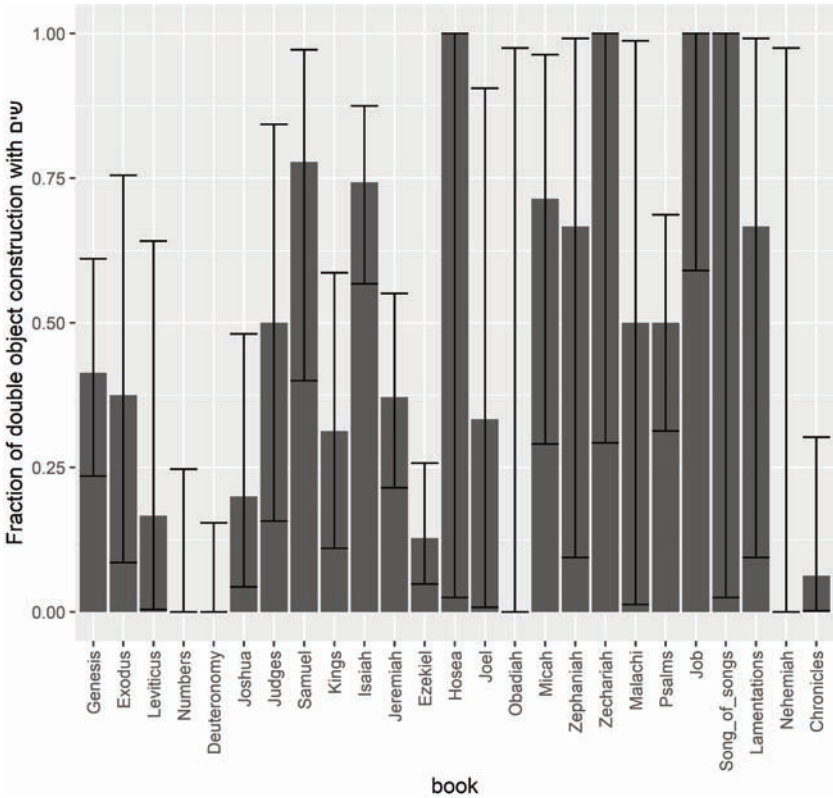

FIGURE 5.11 *Double object constructions with the verb* שים*, as a fraction of the total number of double object constructions of* נתן *and* שים *together*

which do not contain many cases of נתן and שים anyway. For instance, double object constructions with נתן and שים are absent in the LBH books of Esther, Daniel, and Ezra, but Chronicles, the largest of the LBH books, looks more similar to the EBH books.

The two figures show that for both verbs the attestation of double object verbs is relatively high in the books of Isaiah, Jeremiah and Ezekiel, although Ezekiel seems to have a preference for נתן in double object constructions. Also, clearly visible is the contrast between the use of the verbs in Numbers and Deuteronomy. This contrast is less visible in the book of Leviticus due to the low overall attestation of שים in the book of Leviticus.

In all the poetic books, double object constructions with שים are abundantly present, while נתן occurs only in Psalms and Lamentations.

What is the preference of individual books with respect to the two verbs? This becomes clear in figure 5.11 (see page 139). In this figure the fraction of double object constructions is found occurring with the verb שים as a fraction of the total amount of double object constructions occurring with the verbs נתן and שים.

Most books have a mixed profile. They contain double object constructions with both נתן and שים. One reason why a book uses both נתן and שים in syntactically and semantically similar constructions can be that different sources of a book use different verbs. One clear example of layered literature are the books of the Pentateuch, according to the traditional Documentary Hypothesis. In the following excursus, the distribution of double object constructions with נתן and שים is explored in P and non-P parts of the Pentateuch.

5.4.2. *Excursus: Double object constructions of* נתן *and* שים *in the Pentateuch*

The classical Documentary Hypothesis distinguishes four main layers in the Pentateuch: J, E, P, and D.[33] J, E, and P can be found throughout Genesis, Exodus, Leviticus, and Numbers, and D covers the book of Deuteronomy. These redactional layers are subject of extensive discussions, and there are various ways in which the layers can be subdivided. The clearest stylistic differences in the first four books of the Pentateuch can be found between J and E on the one hand, and P on the other.

In the study of diachronic variation, often the redactional layers of the Pentateuch are considered. Wright (2005) is a major publication on the language and date of the Yahwist (J), and particularly P is discussed extensively by various scholars (see also section 2.4). Even though these layers are generally seen as typical representatives of EBH, they each have their own characteristic language.

Double object constructions with נתן can be found 30 times in Genesis–Numbers.[34] Though this construction does not occur often in the first four books in the Hebrew Bible, there is a clear pattern in relation to the different sources. נתן + double object is used predominantly in P, and שים + double object is used predominantly in J and

33 A modern approach to the Documentary Hypothesis is Baden (2012). Important classical works on this topic are the influential *Prolegomena zur Geschichte Israels* by Wellhausen (1883), Driver (1892b), and Eissfeldt (1964). Despite the fact that there is doubt about the independent existence of the sources, especially in European scholarship, the linguistic difference between P, on the one hand, and J and E on the other hand is recognized widely in studies on linguistic variation in BH. See also: Dozeman and Schmid (2006).

34 The letter (J, E, or P) after the verse indicates the source according to Driver (1892b).

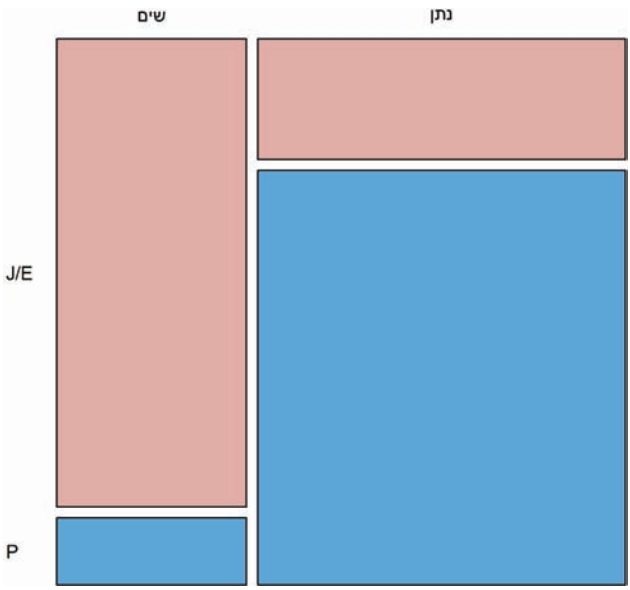

FIGURE 5.12 *Distribution of* נתן *and* שים *with double object constructions in* J/E *and* P

E (non-P).[35] See figure 5.12. Finally, there is Deuteronomy, or the layer D, which can be subdivided in various sublayers. D has a strong preference for using נתן in double object constructions with and without ל introducing the second object. In these constructions, נתן is used 22 times, and שים does not occur.

5.4.3. *Discussion*

It is clear already that in Genesis-Numbers there seems to be an association between the sources of these books and the use of one of the two verbs. Several books, such as Deuteronomy, have a strong preference for נתן. Double object constructions with נתן

35 שים with double object can be found 14 times in J and E and 2 times in P, while נתן with double object is found 23 times in P and 7 times in J and E. All cases can be found in Appendix F. There are some complications related to these numbers. In the analysis, I have relied on Driver's division of the sources of the Pentateuch, and in several cases his division differs from that of others. A case in point is Gen 29:29, which is considered P by Driver, but J by Eissfeldt and Noth (1960). On the other hand, Gen 30:9 is P for Driver, whereas Eissfeldt hesitates between P and J. In other words, if a different division was used the results could be slightly different, although the general picture would be the same.

and שים are rare in most of the core LBH books, and even lacking in Esther, Daniel and Ezra. In Chronicles, a double object construction with שים can be found only once,[36] whereas נתן is used 15 times[37], so it has a strong preference for the verb נתן in double object constructions. This observation is supported by the observations in section 5.3.2, that in four parallel clauses Chronicles uses a double object construction with נתן, where the parallel in Samuel or Kings uses a different verb or construction. Ezekiel is also a book with a strong preference for using נתן, although this book contains several clauses with the verb שים in a double object construction.[38]

Books with a preference for שים are Samuel and Isaiah. In Judges and Psalms, there is an equal use of these verbs in double object constructions and Job uses שים exclusively, although it is more difficult here to speak of a preference, because there is a restricted number of attestations of these double object verbs in the book of Job.

The book of Kings is more mixed, there are some cases in which נתן is used,[39] but there are also cases in which שים occurs.[40] The book of Samuel is mixed as well, with a general preference for שים. Isaiah also has a preference for using שים, but it also uses נתן frequently.

All in all, the evidence for the use of double object constructions with the verbs נתן and שים, shows that there is a preference for the use of one of these constructions in general in poetic and prophetic books, especially in the books of Isaiah, Jeremiah, and Ezekiel. Most books use both נתן and שים, but there are individual books and textual layers that have a strong preference for one of the verbs. Because of the limited amount of data in most books, conclusions need to be drawn with care, but the variation in the choice of the verb found here does not seem to be related to a diachronic development. With the exception of the book of Chronicles, double object constructions with נתן and שים are rare in the LBH books, so it is tentative to suggest a gradual decrease of the use of double object constructions, but more research on, for instance, the DSS and Rabbinic texts is needed to say anything with certainty about this.

---

36 1 Chr 26:10, וישׂימהו אביהו לראשׁ, "His father made him chief".

37 With two direct objects: 1 Chr 16:4, 16:18, 2 Chr 2:10. With a direct object and a ל-object: 1 Chr 17:22, 2 Chr 2:10, 6:27, 7:20, 8:9, 9:8 (2x), 25:16, 25:18, 29:8, 30:7, 35:25.

38 Ezek 7:20, 14:8, 17:5, 19:5, 21:32, 35:4.

39 With two direct objects: 1 Kgs 9:16, 9:22, 11:19, 14:7, 14:9. With a direct object and a ל-object: 1 Kgs 2:17, 2:21, 8:36, 8:50, 2 Kgs 14:9. 1 Kgs 11:19 is a rare case of נתן with two direct objects with the meaning "make her a wife": ויתן־לו אשׁה את־אחות אשׁתו אחות תחפניס הגבירה, "He made for him as a wife the sister of his wife, the sister of his wife Tahpenes".

40 With two direct objects: 1 Kgs 5:23, 2 Kgs 10:8, 17:34. With a direct object and a ל-object: 1 Kgs 10:9, 2 Kgs 10:27.

5.4.4. שים/נתן + *direct object* + ב-*object*

The construction of שים/נתן + direct object + ב-object is rarer than the previously discussed double object constructions with these verbs. With נתן and שים, the construction occurs, respectively, 22 and 27 times in the MT. The construction with ב-object occurs zero or one time in most books, so it is difficult to draw clear conclusions from these data. The meaning of this construction is "to make X like Y".

Some examples are:

> Isa 41:15 וגבעות כמץ תשים
> And you shall make the hills like chaff.

> Hos 2:5 ושׂמתיה כמדבר
> And I make her like a wilderness.

> Song 8:6 שׂימני כחותם על־לבך כחותם על־זרועך
> Make me like a seal upon your heart, as a seal upon your arm.

Figure 5.13 (see next page) shows the counts of this construction per verb in each biblical book.

These 49 cases do not offer much evidence for reaching strong conclusions, but the first impression is that this construction occurs in texts in which other double object constructions occur as well. There is a relatively high frequency in the Major Prophets, especially in Isaiah and Ezekiel. These books have a preference for נתן, just like Chronicles.

Also, for instance, the three occurrences in Genesis with שים can be found outside of the P parts.

> Gen 13:16 J ושׂמתי את־זרעך כעפר הארץ
> I will make your seed like the dust of the earth.

> Gen 32:13 J ושׂמתי את־זרעך כחול הים
> And I will make your seed like the sand of the sea.

> Gen 48:20 E וישׂם את־אפרים לפני מנשׁה
> God makes you like Ephraim and like Manasseh.

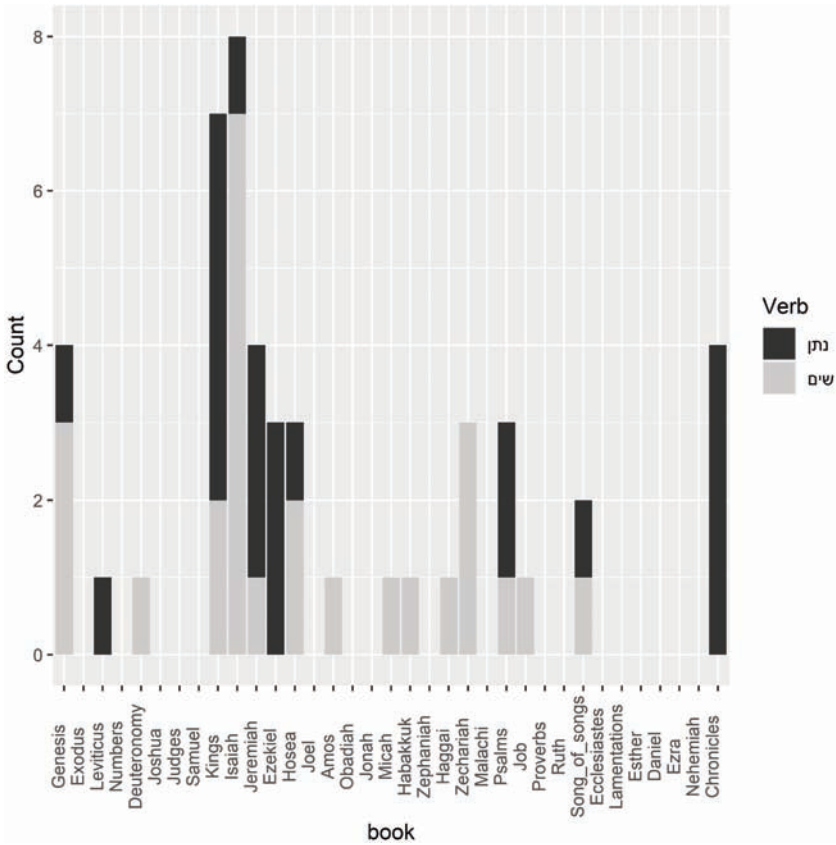

FIGURE 5.13 *Counts of* נתן *and* שים *with double object constructions with* ב*-object*

The construction with נתן occurs twice in the Pentateuch, both times in a P section.

> Gen 9:3 כירק עשֵׂב נתתי לכם את־כל
> Like green herbs I have made it all for you.

> Lev 26:19 ונתתי את־שְׁמיכם כברזל ואת־ארצכם כנחשׁה
> And I will make your sky like iron and your earth like copper.

Also, the construction with שים can be found in the books of Hosea, Micah, Psalms, Job and the Song of Songs, which use predominantly or exclusively שים with one of the other double object constructions.

The construction of נתן with a ב-object can also be found in parallel verses. The following parallel in 1 Kgs 10:27 // 2 Chr 9:27 contains the construction twice.

> ויתן המלך את־הכסף בירושלם כאבנים ואת הארזים נתן כשקמים אשר־בשפלה לרב
>
> The king made silver in Jerusalem as stones, and he made cedars as numerous as the sycamores of the Shephelah.

The figure also shows relatively high frequencies in the Major Prophets of the use of the ב-object in Isaiah, Jeremiah, and Ezekiel, with a preference for נתן in Ezekiel, and a preference for שים in Isaiah.

5.5. Conclusions

This chapter is an exploration of double object constructions of the verbs נתן and שים with an emphasis on cases with two direct objects or a direct object + ל-object. The analysis starts from the perspective that there is a relationship between the valence of the verb and its meaning. This perspective is expressed in previous works by various scholars working with the ETCBC data, especially Dyk, Glanz, and Oosting, but based on the same dataset, a different perspective is definitely possible, such as given by Winther-Nielsen (2017).

Various grammars and lexicons note that the meanings of נתן and שים in double object constructions are closely related. This is an important starting point for this study, because the double object constructions of these verbs are treated as alternatives of the same variable. This semantic similarity is supported by various observations. First, נתן and שים with double object constructions can be found in similar clauses. Second, there is a parallel passage in which Chronicles uses נתן with a double object construction, whereas Kings uses שים (2 Chr 9:8 // 1 Kgs 10:9).

Furthermore, there seems to be no semantic variation between constructions with and without ל introducing the second object. The main reason for this is that constructions with and without ל occur with more or less the same words in similar contexts.

Having said this, there are still various complications. In the double object construction with נתן and שים, there are numerous things that can vary (the verb, the second object can be with or without ל, and there can still be other phrases in the clause influencing the meaning of the verb). An additional complication is

that in most books only a limited number of attestations of these constructions can be found. If in a certain book there is one double object construction with נתן and zero with שים, it does not really make sense to say that the book has a preference for using the verb נתן, because this may well be a coincidental case. Also, there can be more interpretations of certain constructions. Instead of seeing a case like נתן x לאשה–where x is a woman's name, such as in Josh 15:17—as a double object construction, it can also be interpreted as a single object construction with an adjunct. In this research, I have decided to count these cases as double object constructions.

There is a number of double object constructions with the verbs under consideration in parallel texts in the Hebrew Bible. These are mainly parallels between Samuel/Kings and Chronicles. In most of the parallels, the double object constructions are identical to each other, but in a few cases, they are not. There is a case in which Chronicles uses נתן, where שים is used by Kings,[41] there is a case in which Chronicles uses נתן, where כון is used by Samuel,[42] there is a case in which Chronicles has a double object construction, where Kings has only a single object,[43] and, finally, there is a case where היה is used Kings.[44] In these cases, Chronicles uses נתן with a double object construction, where Samuel and Kings have an alternative. Chronicles has a strong preference for using נתן with a double object construction, whereas in other books the image is more mixed.

How are נתן and שים with double object constructions distributed throughout the different levels of the main variables of this research? There is not much difference between the choice for נתן or שים between Q and N. Also, between main and subordinate clauses there is no clear preference for one of the two verbs. Both verbs occur with double object constructions in main clauses, adverbial and relative clauses, but not in argument clauses. Comparing EBH and LBH, it is striking that נתן and שים are more or less absent in Esther, Daniel, Ezra, and Nehemiah, which makes it impossible to draw conclusions about these books. In the book of Chronicles, however, double object constructions occur abundantly with נתן, and only once with שים. The

41 1 Kgs 10:9 // 2 Chr 9:8 ויתנך עליהם למלך לעשות משפט וצדקה // וישימך למלך לעשות משפט וצדקה.

42 2 Sam 7:24 // 1 Chr 17:22 ותתן את־עמך ישׂראל לך לעם עד־עולם // ותכונן לך את־עמך ישׂראל לך לעם עד־עולם.

43 1 Kgs 10:9 // 2 Chr 9:8 לתתך על־כסאו למלך ליהוה אלהיך // לתתך על־כסא ישׂראל.

44 1 Kgs 9:7 // 2 Chr 7:20 ואתננו למשׁל ולשׁנינה בכל־העמים // והיה ישׂראל למשׁל ולשׁנינה בכל־העמים.

impression that Chronicles has a strong preference for using נתן is strengthened by the observation of four parallel clauses in which Chronicles uses נתן with a double object construction, whereas the EBH alternative uses a different verb or נתן with a different construction. In the Pentateuch, P and D have a strong preference for using נתן, but J and E predominantly use שים. Other books and texts seem to have a mixed profile, or not enough data to draw conclusions from.

There is a general increased use of both verbs in prophecy, especially in the Major Prophets. Of these three books, Ezekiel has a strong preference for נתן, whereas Isaiah prefers שים. Jeremiah has a slight preference for נתן.

The use of נתן and שים with two objects, of which the second object is introduced by ב, is rare, occuring only 49 times in the MT. In the individual biblical books, it generally occurs only a few times. The main impression of its distribution is similar to the constructions already discussed. It occurs with a higher frequency in the Major Prophets, Chronicles has a strong preference for the verb נתן, just like P, and J and E prefer to use שים.

The variation between נתן and שים with two objects seems to have a varied background. On the one hand, both verbs occur often with these constructions in the Major Prophets, on the other hand, some books have a strong preference for one of the alternative verbs. There might be a preference for נתן in exilic/post-exilic Hebrew. Chronicles has a strong preference for this verb, and the same is true for P (if it is accepted that P is late) and Ezekiel.

Even though both verbs with double object constructions are relatively frequent in BH, there remain various questions. How do these constructions behave in postbiblical Hebrew, and how does that relate to BH? Other verbs than נתן and שים are also synonymous. Do patterns of those verbs cohere with the patterns found here? This research is based on the idea, that נתן and שים with double object constructions are synonymous. Based on various observations this seems reasonable, but there is a chance that there are cases of idiomatic expressions, in which only one of the alternatives can be used. It is difficult to find evidence for this, especially if such expressions are rare, but maybe further research can shed light on this issue.

All in all, there are some interesting tendencies visible. Of the main variables of this research, main and subordinate clauses and discourse type do not play a significant role. Genre is interesting, because there is an increased use of double object patterns with נתן and שים, but this is mainly visible in the Major Prophets. The most interesting variation can be observed between individual books and redactional layers of books. Some books or layers have a strong preference for one of the two

alternative verbs in double object constructions, and even though the number of attestations in most books is relatively low, some preliminary conclusions could be drawn, as was shown above. Research on other verbs and extrabiblical texts can shed more light on the patterns observed here.

CHAPTER 6

Clause structure variation using sequence analysis

6.1. Introduction

The previous chapters focused on specific linguistic features and how the use of linguistic alternatives is conditioned by the linguistic and non-linguistic environment. It is also possible to study language in a way in which not only a small selection of clauses containing a specific feature is analyzed, but the majority or all clauses from subcorpora. Depending on how clauses are represented, linguistic variation can be modeled using a large number of linguistic features in the selected clauses. An advantage of such an approach is that it gives a far more complete image of linguistic variation than studies of individual features.

A topic that may profit from an approach like this, is the question as to what extent EBH and LBH can be distinguished from one another. In most studies on this issue, linguistic features are extracted from the EBH and LBH texts following the traditional procedure, as described in section 2.3. An unsolved problem here is to what extent EBH and LBH differ. It is clear that there is linguistic variation in BH, and also that there is linguistic variation between EBH and LBH, but if a clause is selected randomly from the EBH and LBH books, it is hard to say whether the clause is selected from EBH or LBH, because most clauses are structured in a way that is common to both EBH and LBH.

It is equally difficult to say whether the syntax of a clause from one of the books of disputed date is more characteristic of EBH or LBH. The books of Ruth and Jonah and the prose tale of Job have been discussed extensively in the literature on diachrony in BH.[1] Hurvitz (1974) argued that the prose tale of Job is written in LBH on the basis of seven late linguistic features in it. Young (2009) disagrees with Hurvitz on this conclusion. He argues that it is true that some LBH features can be found in the prose tale of Job, but the accumulation of these features is lower than in core LBH texts, the density is even comparable with that of various EBH texts. The problem of the low density of LBH features is a general problem concerning the difference

1 An overview of the literature on the books of Jonah, Job, and Ruth can be found respectively on the pages 43–45, 53–56, and 58–60 of Young, Rezetko, and Ehrensvärd (2008, volume 2).

between EBH and LBH, because there is no clearly defined lower limit of LBH features for LBH texts. Similar discussion of the books of Jonah and Ruth can be found. Young, Rezetko, and Ehrensvärd (2008, volume 2: 43–45) refer to various scholars who date the language of Jonah as late, based on the attestation of Aramaisms and features that can be found elsewhere in LBH books. On the other hand, there are others who argue that its language shows dialectal traces, and Landes (1992: 130) argues that the linguistic evidence is not conclusive for dating the book precisely. Young, Rezetko, and Ehrensvärd (2008, volume 2: 58–60) see a tendency in recent scholarship to date the book of Ruth as post-exilic, based on linguistic and non-linguistic arguments. They also note that there is a variety of judgements on the character of its language, but most scholars seem to agree that it resembles mainly EBH.

In this chapter, the problem of the difference between EBH and LBH is approached from a different angle. Instead of only selecting clauses with specific characteristics, the majority of clauses from EBH and LBH will be taken into account to find out to what extent EBH and LBH can be distinguished. The following questions will be answered:

> To what extent do EBH and LBH differ with respect to clause structure? Does the clause structure of clauses in the books of Ruth, Jonah, and the prose tale of Job mainly reflect the characteristics of EBH or those of LBH?

The general approach is as follows. A sample of Q and N clauses are taken from the EBH and LBH corpora. Then a model is trained on the Q clauses and another model is trained on the N clauses, with the language phase as target variable. Next, the model will be used to predict the "class", which is either EBH or LBH, of clauses from the books of Jonah and Ruth and the prose tale of Job. Also, the predictions of the language class are made for one EBH or LBH book, which was kept separate during the training phase.

The linguistic features on which the model will be trained can be selected and structured in a wide variety of ways. In this research, the model is based as much as possible on syntactic structure instead of on the lexicon. Therefore, the analysis is done with clauses structured in two ways, as a sequence of phrase functions and as a sequence of parts of speech. Only the basic skeleton of the language remains. This means that many clauses are similar to each other and that there is little variation is to be expected between EBH and LBH. It is likely that the accuracy of the predictions on the test set will be only slightly higher than 50%.

The tool that will be used to train the model is an LSTM network, which can be used for modeling all kinds of sequence phenomena. Neural networks can give slightly different results on the basis of the initialization of the weights and variation in the training set due to sampling. This is an issue for the present dataset, consisting of clauses from EBH and LBH. As stated, there is not much variation to be expected between these subcorpora, and the amount of data is limited, so a bit of variation in the results caused by variation in the initialization of parameters can seriously distort the analysis. Therefore, not just one model will be trained for Q and N, but 200 independent models, and the results will be averaged to stablilize the results. A model based on many separate models together is called ensemble learning, which is also used in, for instance, the Random Forest.

If the accuracy of the predictions is only slightly higher than 50%, how can we know that this approach works? To check this, an extra analysis will be done as a kind of validation of the whole approach. In this analysis, the dataset consists of clauses in BH and Biblical Aramaic (BA). The rest of the analysis is similar to that of the variation between EBH and LBH. If the LSTM model is able to distinguish between Hebrew and Aramaic, it is to be expected that the prediction accuracy on the test set is substantially higher than in the case of the overall prediction accuracy for EBH and LBH.

Just like in the case of the analyses of sections 4.3 and 4.4, this problem is designed as a classification problem. Instead of using the supposed language phase as predictor, in this analysis, it is the output variable.

6.2. Analyzing sequence data

6.2.1. *Introduction*

Sequence data, like language, can be analyzed in various ways. A traditional way of modeling sequences is by using n-grams. With such an approach subsequences of length n are counted in a text or corpus of texts, and the resulting dataset can be analyzed with algorithms that can be used for the analysis of structured data. Another traditional way of modeling a sequence is by using a Markov Model, which is explained in the next section (6.2.2). In this research, a Recurrent Neural Network (RNN) is used to model clauses. This is a more modern and flexible approach to sequence modeling. The RNN is introduced in section 6.3.3.

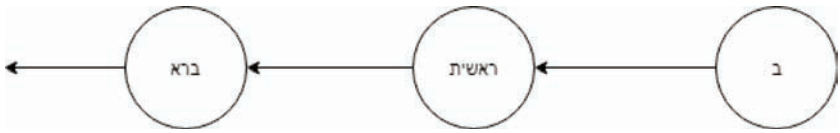

FIGURE 6.1 *A Markov chain*

6.2.2. *Markov Chains*

One way of modeling sequences is by using a Markov chain, which was the approach chosen for the same problem as in the research by Van de Bijl, Kingham, Van Peursen, and Bhulai (2018). A Markov chain models the dynamics of a system which changes over time with regular time steps. A Markov model based on texts can be constructed with a word on every time step, but other choices are possible as well, such as a phrase or a clause. The Markov Model is based on the Markov assumption, which states that the present state is dependent only on the state in the previous time step and not on what happened earlier in time. In texts, a state can simply be a word and the whole sequence is the clause in which the words occur. See figure 6.1.

In the Markov Model, for every transition from one state to another the probability of this transition is calculated, and then, the probability of finding the whole sequence can be calculated.

In figure 6.1, the state ראשית only depends on the state ב, and the state ברא only depends on the state ראשית. This simplification is useful for the calculation and probably also often works well in practice, but in language there are long term dependencies. One way to solve this is to work with n-grams. In the case of the sequence in figure 6.1, if n is 2, the first state is ב-ראשית, the second state is ראשית-ברא, and so on. This only partly solves the problem of the long-term dependencies, because there may be dependencies at a longer distance than just a few words. An additional problem of the n-grams is that if n becomes larger, the individual n-grams become less frequent, which makes it more difficult to generalize the results.

An extension of the Markov Model is the Hidden Markov Model (HMM), which is based on the concept of hidden states. It can be used for applications like part of speech tagging. In the HMM the state is accompanied by a hidden state. In the clause בראשׁית ברא אלהים את השׁמים ואת הארץ, the states are the concrete words while the hidden states are the parts of speech of these words (see figure 6.2), although the hidden state can be any characteristic of the word, like gender or number.

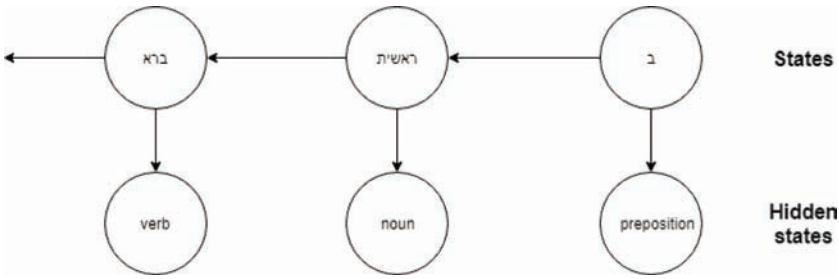

FIGURE 6.2 *A Hidden Markov Model*

6.2.3. *Neural Networks*

An Artificial Neural Network, or simply Neural Network (NN) is a mathematical structure, which owes its name to the biological equivalent, because just like in the biological network, information is passed through series of neurons. With NNs, it is possible to study patterns in data that are strongly non-linear, such as texts or pictures. In recent years, there has been much research on NNs, which has led to a variety of different architectures of networks. These have made NNs useful for many practical applications related to Natural Language Processing and computer vision, which can be used simultaneously in applications like self-driving cars.

The NN consists of layers of neurons. If the number of layers in the network is high, the network is called a deep network. There is no clear definition of how many layers a network needs to have to be called a deep network.

A special kind of architecture of NNs is the Recurrent Neural Network (RNN). The RNN is used for analyzing numeric and non-numeric sequences and it has the ability to "remember" what has happened earlier in the sequence of events. Several improvements of the ordinary RNN have been proposed, which solve certain problems during the training of an RNN-model. One of these new developments is the Long Short-Term Memory (LSTM) network, which is used in this research.

Various frameworks for working with neural networks have been published since 2015. Among the most important ones are TensorFlow and Keras. TensorFlow[2] is created and maintained by Google and was made open source in 2015. In this research, Keras[3] is used with a TensorFlow backend. Keras is a wrapper for Tensorflow and other frameworks and can be used for fast experimentation.

2 www.tensorflow.org.
3 https://keras.io.

Recurrent neural nets are used for analyzing all kinds of sequential data. These sequences can be numeric time series like stock indexes or average daily temperatures throughout the year, but an important area of applications is in the field of Natural Language Processing (NLP). Applications in which LSTM models can be used are chat bots, in which a computer responds to a human question in natural language, and translation machines, but they can also be used to generate creative text, like poetry.

The way the LSTM network learns from the features differs substantially from the way Random Forest or XGBoost learns. In the case of those algorithms, the supposedly relevant features should be selected and organized in a structured dataset. In the case of sequence classification with LSTM, the sequences only need to be given to the network, and the network finds out itself which features are relevant. These features can be complicated, and a disadvantage of using this approach is that the researcher has no way to access which features are important for classification of the clauses. A good reason for still using LSTM networks for sequence analysis is that they often offer the best results from the perspective of accuracy.

More information about NNs and LSTM networks can be found in Appendix E.

6.2.4. *Experimental design*

The model used in this research consists of an embedding layer, followed by two LSTM layers. It has a final dense layer of one neuron, which is characteristic of neural nets with a binary output (EBH and LBH in this case). Both LSTM layers have a 0.5 dropout rate to avoid overfitting.

The data are divided in two groups of clauses:

- Q clauses (EBH and LBH)
- N clauses (EBH and LBH)

A random sample will be taken from both Q or N groups, in such a way that a sample contains as many clauses from EBH as from LBH.[4] Also, clauses from one book from the EBH and LBH subcorpora are held separate.

Next, models are trained on the Q clauses and N clauses separately, and predictions are made using this model on the test set, which are the Q and N clauses from one book that was not included in the training set. This procedure of using a separate

4 This is actually a case of downsampling, in contrast to upsampling, which was used in chapter 4.

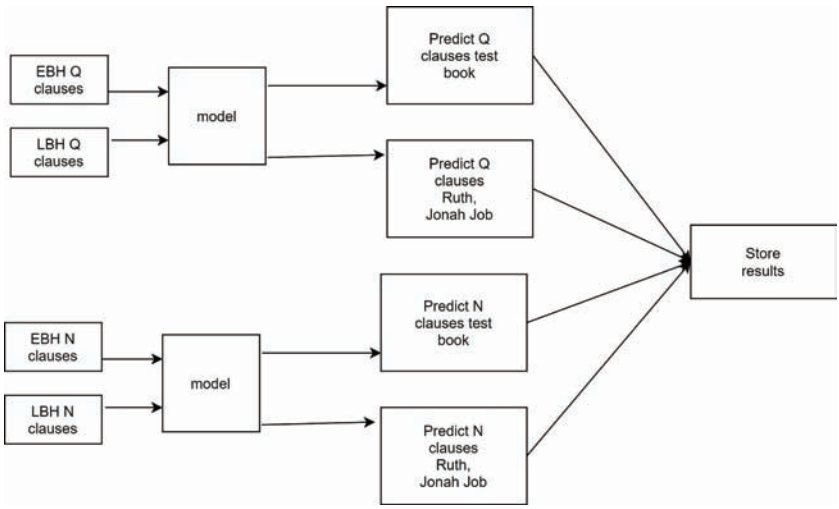

FIGURE 6.3 *Design of the analysis*

book as test set instead of sampling randomly from all the Q or N clauses is chosen, in order to avoid the presence of clauses from one book in both the training and test set, which can be seen as a kind of contamination. See figure 6.3 for an overview.

The procedure in this figure is repeated 200 times for each book as test set. The difference between EBH and LBH is relatively small, and due to variation in the initialization of parameters and sampling, there can be some variation in the results.[5] By averaging the results of many models, the results are stabilized.

Next to the predictions made on the test book, predictions are made on Q and N clauses from the books of Jonah, Ruth and the prose tale of Job.

The result of these steps is that all the N and Q clauses in the EBH and LBH books and the books of Jonah, Ruth, and the prose tale of Job have been classified multiple times as EBH or LBH. Finally, the results are analyzed using cluster analysis to find out which books share similarities.

5 For the problem of high-variance, see James, et al. (2013: 33–35).

6.3. Data preparation

Each sequence that is processed by the network consists of one clause from the EBH and LBH subcorpora.[6] A clause from the Hebrew Bible can be represented in different ways, for instance, a clause from Gen 1:2 with full vocalization:

וְחֹשֶׁךְ עַל־פְּנֵי תְהֹום

It can also be written without vocalization:

וחשך על־פני תהום

or, in ETCBC transcription:

W XCK <L PNJ THWM

With representations as the ones above, the vocabulary plays an important role in an analysis. This influence can be avoided by representing the clause as a sequence of parts of speech. The same clause looks then as follows:

conj - subs - prep - subs - subs

If the focus of an analysis is on clause syntax, one can represent this clause as a sequence of three phrases and represent it by the types of the phrases:

CP - NP - PP

Another option, also based on representing the clause as a sequence of phrases, is to represent it using phrase functions:

Conjunction - Subject - Predicate Complement

In section 6.3.1 clauses are represented in the latter way, as a sequence of phrase functions.

---

6 The scripts for processing and classifying EBH and LBH data can be found here: https://github.com/MartijnNaaijer/phdthesis/tree/master/Ch6_Sequence_analysis/classify_EBH_LBH.

In section 6.3.2 a similar analysis is done, but with a representation of clauses on the word level using parts of speech.

The texts used to train the model are Q and N clauses from Genesis–Kings as best representatives of EBH and Esther–Chronicles as best representatives of LBH.[7]

6.3.1. *Conversion to numbers*

The data are, as usual, extracted from the ETCBC database using Text-Fabric, but a clause consisting of four phrases, such as Conj - Pred - Subj - Cmpl cannot be processed directly by the network. The first step is to convert the features into numbers, because the network can only handle numeric data. Therefore, every phrase is converted into an integer, the phrase with the highest frequency gets the lowest value, which is 1, which means that the given clause could be converted into:

1 - 2 - 4 - 3

The result is a dataset in which all clauses are represented as arrays of integers. Another preparation step is needed. The clauses contain varying numbers of phrases, but the LSTM network can only deal with sequences of equal lengths. To solve this problem, the sequences are padded with zeros. Suppose that the longest sequence in the training set contains nine phrases, then the clause mentioned before is converted into the following sequence:

0 - 0 - 0 - 0 - 0 - 1 - 2 - 4 - 3

By adding five zeros, the sequence has a length of nine now. The data are processed further into a format called one-hot encoding. Suppose the input data consists of an alphabet of three characters. Then their numeric representation is 1, 2, and 3. In one-hot encoding, this is [1, 0, 0], [0, 1, 0], and [0, 0, 1]. Each character is converted into a list of zeros, each with the length of the alphabet, with a one, of which the index indicates its numeric value. After this step all sequences are ready to be processed by the network.

---

7 Practically, this means that those clauses are selected that have "N" or "Q" as the last character in the "txt" feature in the ETCBC database. This means also that clauses having "D" or "?" as the last character will not be selected, since these classes are less well defined and less frequent than "N" and "Q".

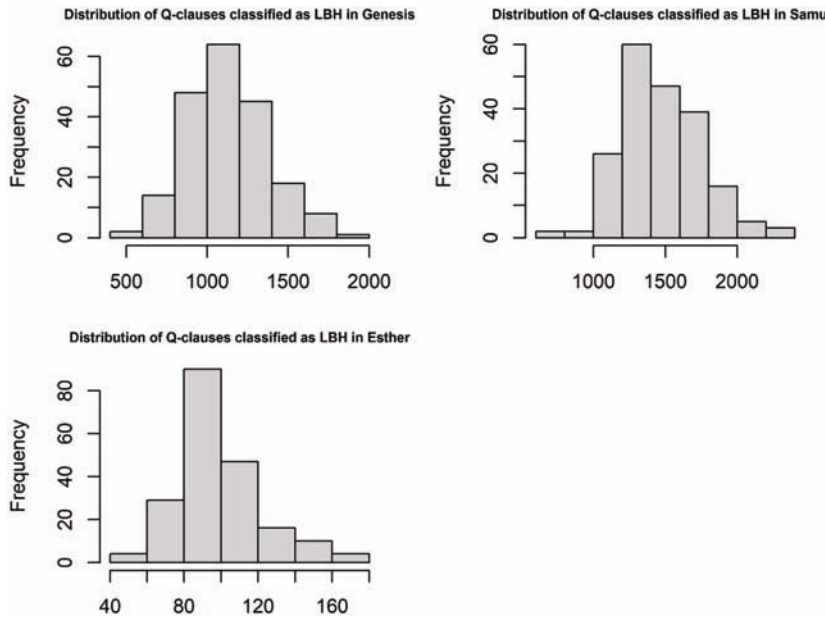

FIGURE 6.4 *Distribution of* Q*-clauses classified as* LBH *for* EBH *and* LBH *books*

6.4. Results

6.4.1. *Results of the phrase level model*

Preliminary exploration of the results

As already said, the predictions between different runs of the algorithm can vary substantially due to variation between samples and the way the weights of the model are initialized. As examples, figures 6.4 and 6.5 (see next page) show the distribution of predictions of the 200 times that the model has been run for the Q-clauses in the books of Genesis, Samuel, and Esther. For Jonah, Ruth and Job all the predictions together are plotted. There are far more predictions for these texts than for the test books from the EBH and LBH subcorpora, because for every test book 200 predictions were made on these texts. Figures 6.6 and 6.7 (see next pages) show similar distributions for the same books for the N-clauses.

The histograms show that the distribution is more or less symmetrical around the mean value, which is what one would expect. They also show that the number of clauses classified as LBH can vary strongly within a book. For instance, in the case of

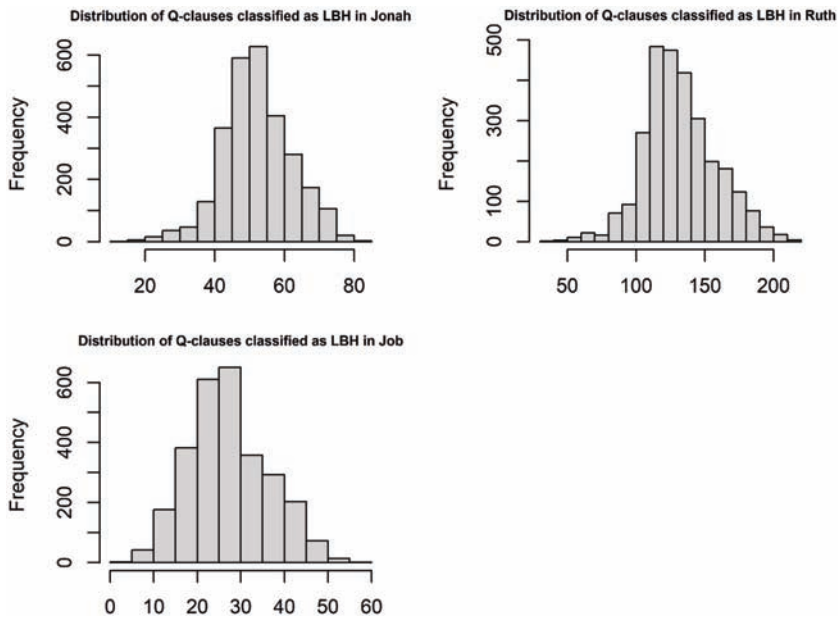

FIGURE 6.5 *Distribution of* Q*-clauses classified as* LBH *for books of uncertain date*

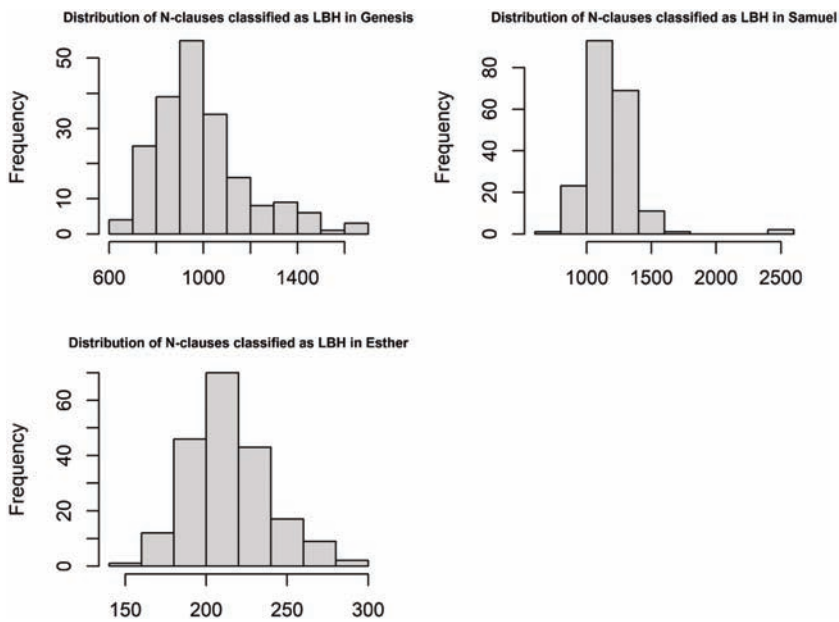

FIGURE 6.6 *Distribution of* N*-clauses classified as* LBH *for* EBH *and* LBH *books*

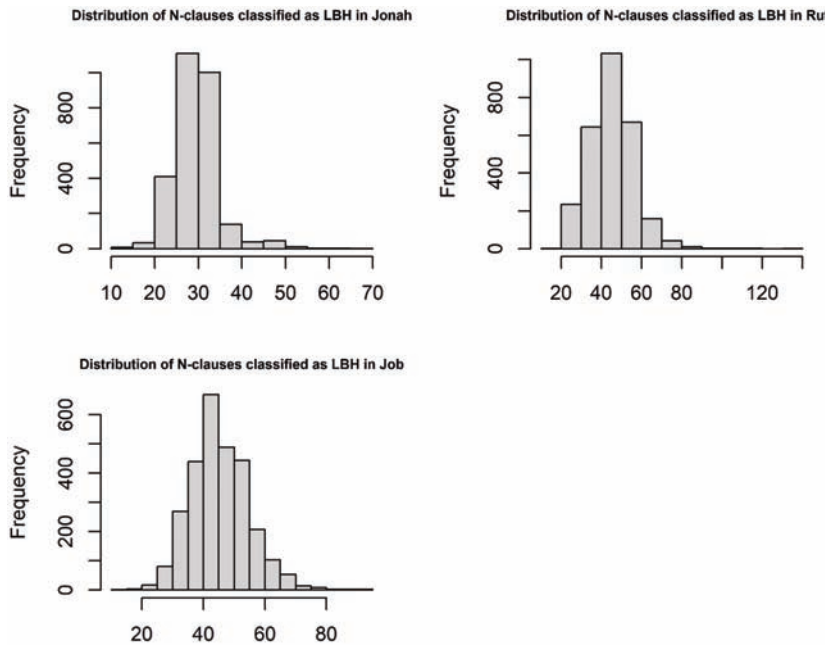

FIGURE 6.7 *Distribution of* N*-clauses classified as* LBH *for books of uncertain date*

the N-clauses in the book of Ruth, in some models fewer than 20 clauses are classified as LBH, whereas in some models more than 100 clauses are classified as LBH. It is this relatively large spread of the results which makes it necessary to work with an ensemble of models.

The mean number of clauses classified as LBH are used for the analysis. An important issue is whether this mean is stable enough. If the mean is calculated on the basis of five models it is, of course, more stable than if it is calculated on the basis of two models, but how many are needed to make the predicted number really stable? To explore the stability of the 200 models for each of the EBH and LBH test books figures 6.8–6.11 (see the following pages) show the cumulative mean of the predictions for the same books as in figures 6.4–6.7.

In these figures, the mean number of clauses classified as LBH is shown after every new prediction. After only a few runs of the model there are strong fluctuations in the average prediction score, but after about 50 runs the average is more or less stable, so 200 runs are clearly enough to produce a stable average.

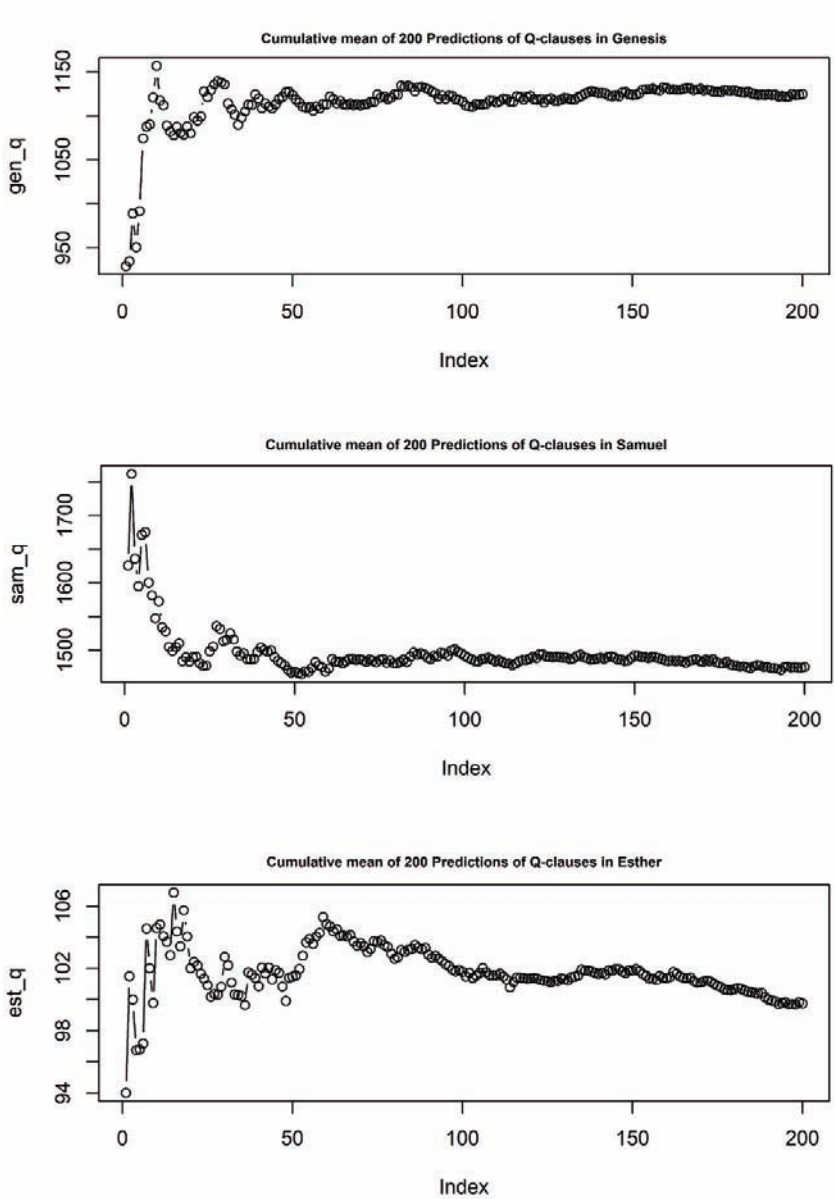

FIGURE 6.8 *Cumulative mean of predictions of* Q*-clauses for* EBH *and* LBH *books*

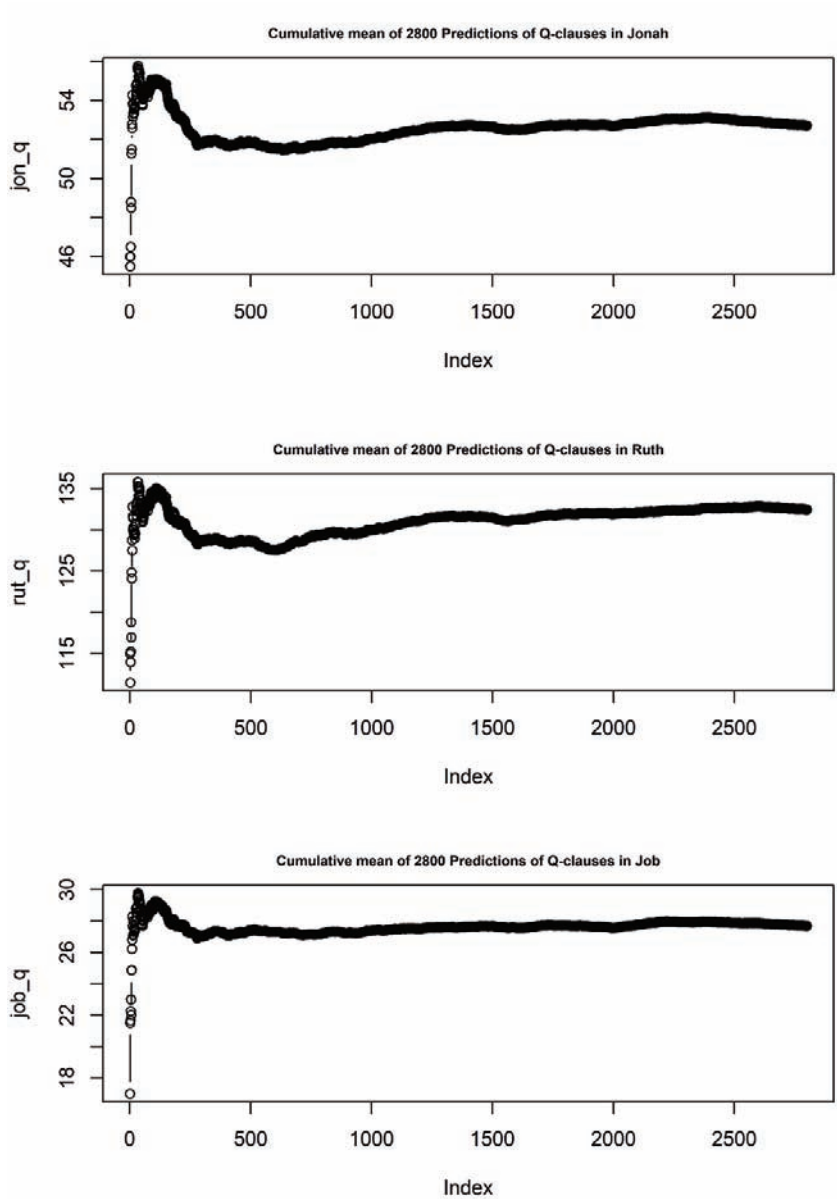

FIGURE 6.9 *Cumulative mean of predictions of* Q*-clauses for books of uncertain date*

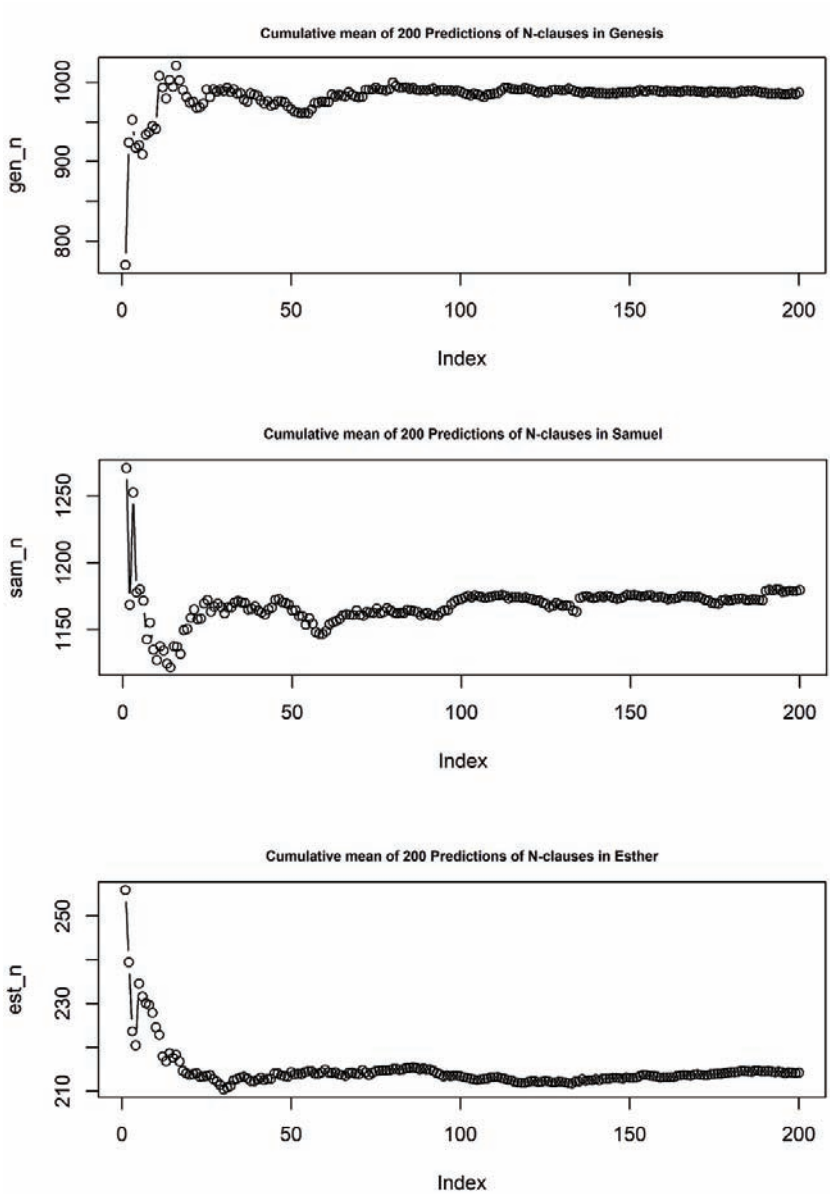

FIGURE 6.10 *Cumulative mean of predictions of* N *clauses for* EBH *and* LBH *books*

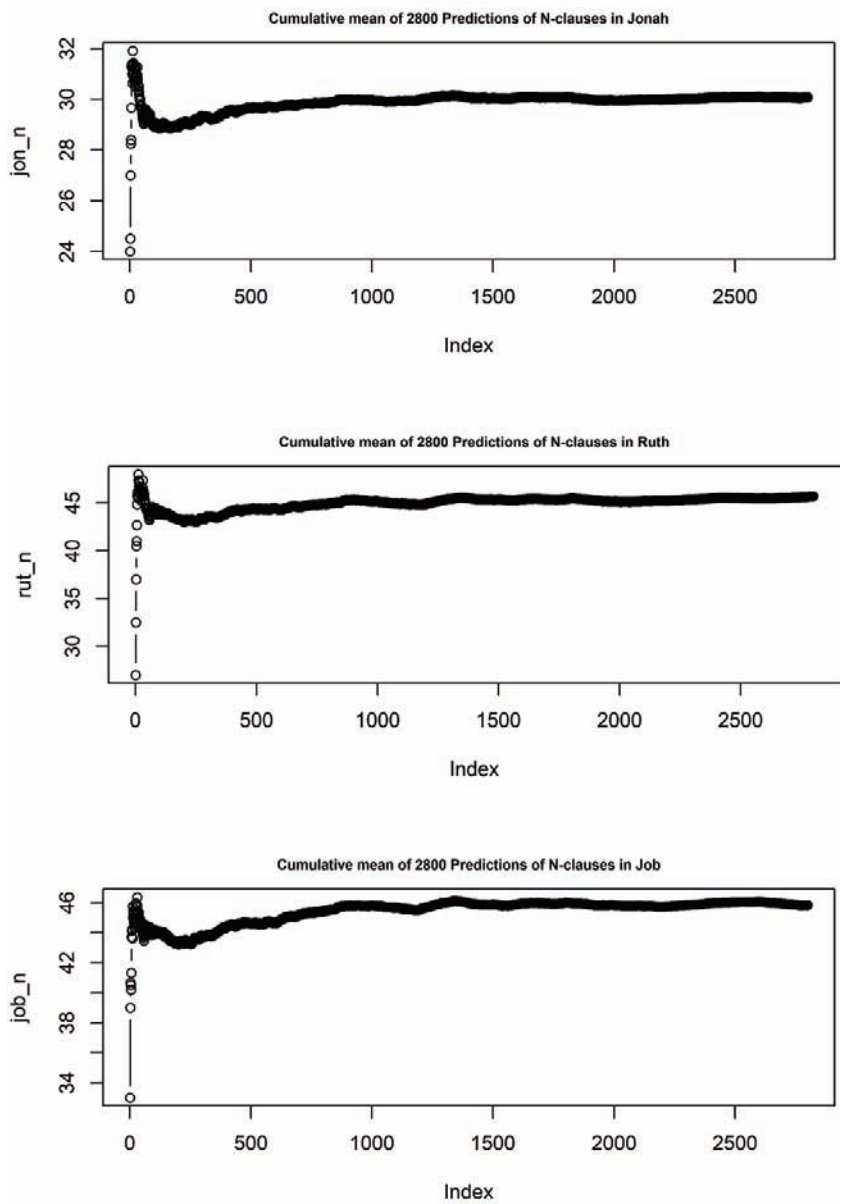

FIGURE 6.11 *Cumulative mean of predictions of* N *clauses for books of uncertain date*

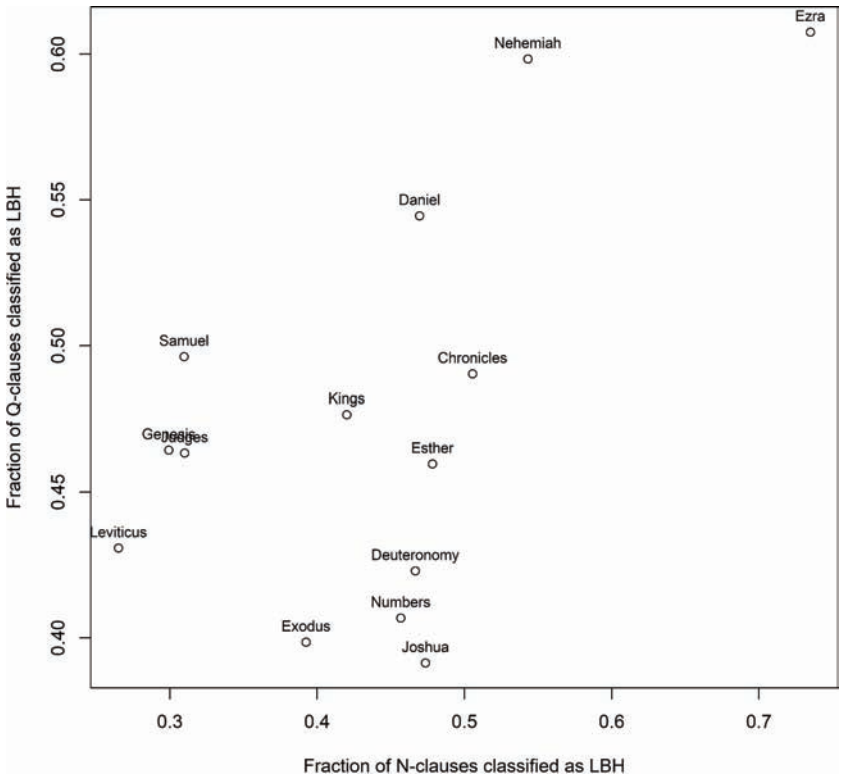

FIGURE 6.12 *Fraction of clauses classified as* LBH *(phrase level)*

Analysis

In the previous explorations, the results were shown as the number of clauses in a book that were classified as LBH by a model, but more important is the fraction of the total amount of Q and N clauses in a book classified as LBH that needs to be taken into account. The expectation is that for both Q and N clauses the LBH books have a higher concentration of clauses classified as LBH. Figure 6.12 shows the results in a scatterplot.

On the vertical axis fraction of Q clauses classified as LBH are plotted and on the horizontal axis these values for the N clauses. On both axes a value of 0.5 indicates that on average 50% of the clauses is predicted in the LBH class and 50% has been predicted in the EBH class. A value of 0 means that all clauses were classified as EBH and a value of 1 indicates that all clauses were classified as LBH. Based on the literature on linguistic variation in BH, one expects EBH books to be found in the lower left corner and LBH books in the upper right corner.

The first interesting thing to observe is that there is more variation on the N axis than on the Q axis. On the Q axis, average values vary roughly between 0.4 and 0.6 and on the N axis there is variation between 0.25 and 0.75. This means, that less variation between EBH and LBH can be found in Q clauses than in N clauses.

The EBH books behave more or less as expected. All EBH books can be found in the lower left corner of the figure, which means that they have a score lower than 0.5 on both axes.

On the N-axis, the books of Esther, Daniel, and Chronicles do not score higher than 0.5, but the scores are higher than most of the EBH books. This means that according to the present approach, most of the language of these books is not shared with LBH, but with EBH. Having said that, the core LBH books behave as expected in the sense that if they are considered as a group, they can be found more in the right and upper part of the figure, although there is substantial variation between the LBH books.

The higest LBH scores are found in Ezra and Nehemiah. One explanation for this result is that Ezra, and to a lesser extent Nehemiah, contain the highest concentration of late language of all the books under consideration, but an alternative explanation is that the language of these two books is relatively close together. This can be because the language of these books is simply relatively late, but it is also possible that the language is similar because these books are often regarded to be one book, Ezra-Nehemiah.

What is the situation in the case of the texts of unknown date, Jonah, Ruth and the prose tale of Job? If we want to classify these texts as EBH or LBH, we can use a classifier, such as the straight line in figure 6.13 (see next page).

In the figure, the classifier is linear and it was drawn manually. The texts right of the line are similar to LBH, and everything left of it is similar to EBH. One can also opt for a non-linear classifier, but with only a relatively small number books as references, non-linear classifiers can easily lead to overfitting. In the given situation, one does not need very sophisticated tools to see that on the N axis Jonah, Ruth, and Job share most of their language with the EBH books, and that on the Q axis Jonah and Ruth score higher than 0.5, but Job scores particulary low here.

A different way of looking at these results is by clustering the books. Figure 6.14 (see page 168) shows the results of k-means clustering. In k-means, one has to specify the number of clusters, which is the k in k-means. Given the underlying presupposition that the data can be divided in two groups, EBH and LBH, two clusters are made.

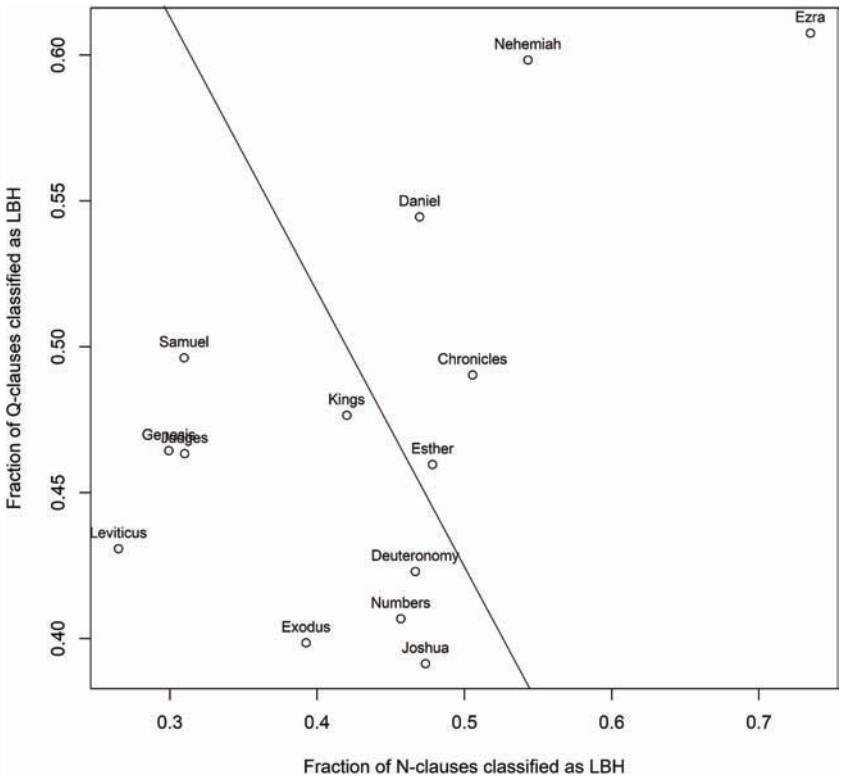

FIGURE 6.13 *Classifying texts of unknown date using a linear classifier*

One cluster contains the books Genesis, Leviticus, Judges, Samuel, Jonah, Ruth, and the prose tale of Job, and the other cluster contains the books of Exodus, Numbers, Deuteronomy, Joshua, Kings, Esther, Daniel and Chronicles.

The clusters in the figure can be found more or less next to each other. They are based mainly on variation on the N-axis, which is as expected, because the variation on this axis represents most of the variation in the dataset. The cluster on the left side of the figure contains only EBH books and the three texts of unknown date, and the right cluster contains all the LBH books and some EBH books.

The clusters show what was more or less clear from the previous observations: in the first place there is no clear distinction between EBH and LBH. There is a tendency in the data that the LBH books fall on one side of the figure and the EBH books fall on the other side, but without any knowledge of the data, it would not be possible to distinguish between EBH and LBH. In the second place, Jonah, Ruth, and the prose tale of Job are grouped in the class containing only EBH books.

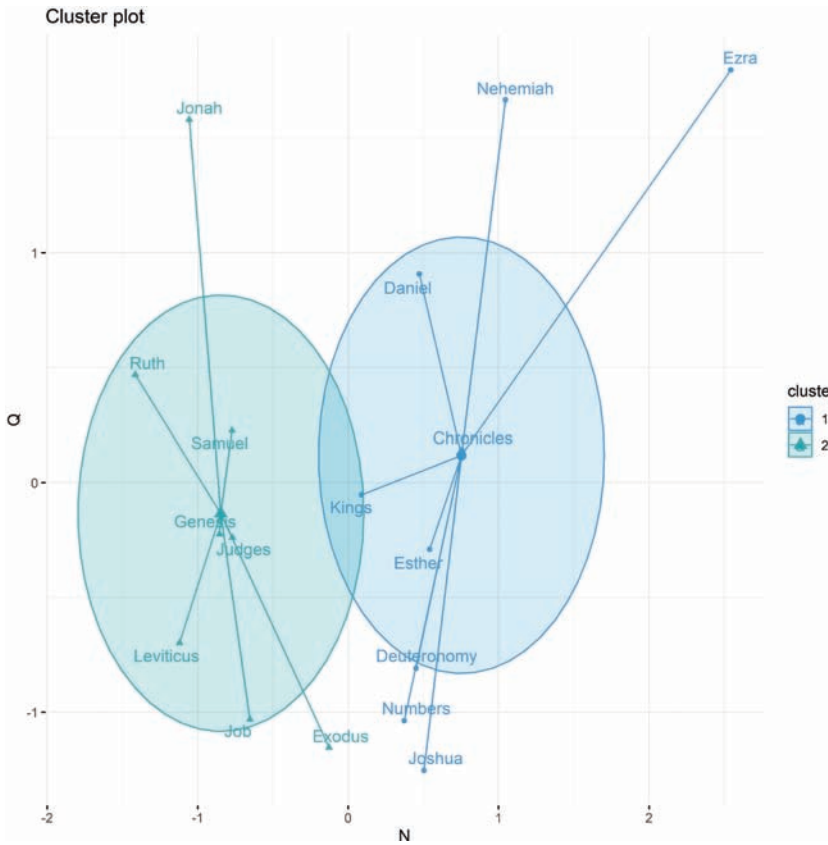

FIGURE 6.14 *Cluster analysis of* EBH *and* LBH *books (phrase level)*

6.4.2. *Results of the word level model*

The only difference between the word level analysis and the phrase level analysis is that the clauses are represented as a sequence of parts of speech instead of as a sequence of phrase types.

Figure 6.15 shows the scatterplot with the average predictions for running the model 200 times for each of the EBH and LBH books.

The figure shows some similarities and some differences in comparison with figure 6.12. As in the phrase level analysis, Ezra and Nehemiah are the main "outliers" in the plot, where the rest of the EBH and LBH books stick relatively close together. This is caused mainly by the low variation on the Q-axis. All the books, with the exception of Nehemiah, have a value between 0.25 and 0.35 on this axis. On the N-axis

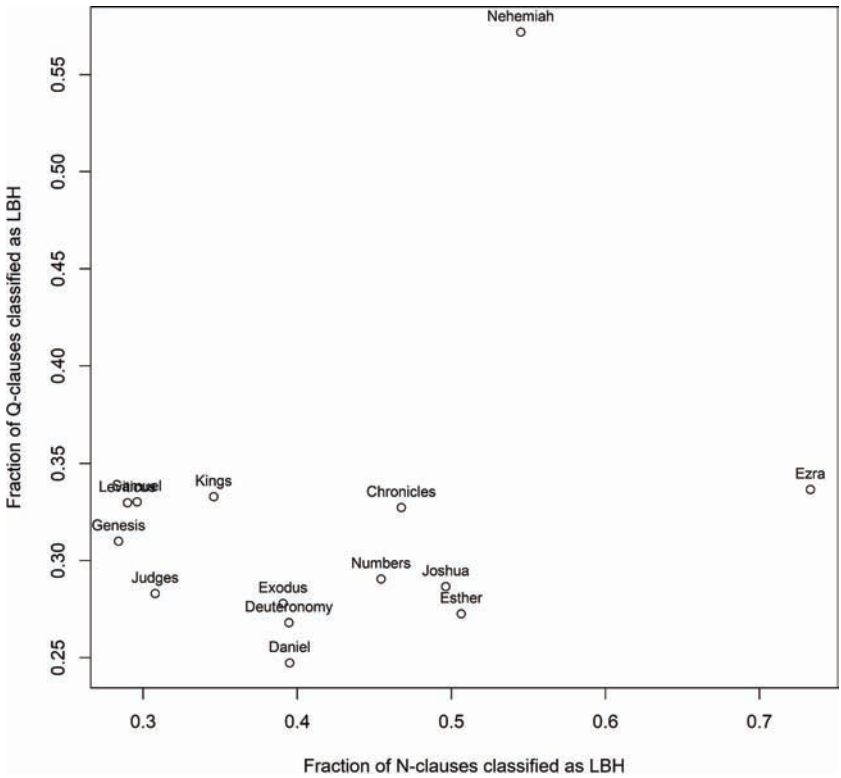

FIGURE 6.15 *Fraction of clauses classified as* LBH *(word level)*

the results are similar to those in the phrase level model. Figure 6.16 (see next page) shows the k-means clusters for the word level analysis, again, with two clusters. The clusters are similar to those in the phrase level analysis in figure 6.14. The left cluster contains the three texts of unknown date, the book of Daniel and most of the EBH books, and the right cluster contains the books of Numbers and Joshua and the other LBH books. Some books in the "border region" of the cluster changed from one cluster to the other, but these are only minor differences.

6.4.3. *Classification of Hebrew and Aramaic clauses*

It is clear that BH is a relatively uniform language: there is no sharp distinction between EBH and LBH. Many clauses that can be found in EBH can also be found in LBH, and this is also literally the case if one looks at clauses occurring in parallel texts in Samuel/Kings and Chronicles. If one wants to classify separate clauses as EBH or

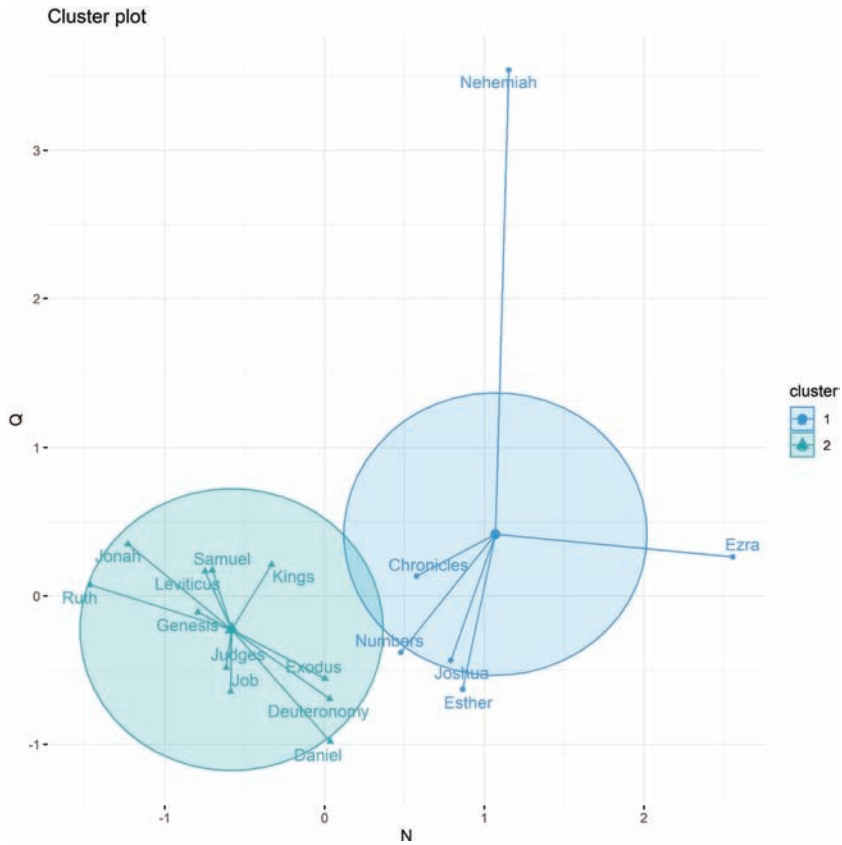

FIGURE 6.16 *Cluster analysis of* EBH *and* LBH *(word level)*

LBH it is to be expected that the correct classification score will not be much higher than the random guess of 0.5. This can also be seen in the learning process of the network. If the LSTM network has to learn the difference between clauses in EBH and LBH, it hardly learns anything. There is only a small decrease of the loss,[8] and an additional step, the ensemble approach, is needed to stabilize the results.

How do we know that such an approach actually works well? Does this method produce clearer results if two distinct languages are chosen, instead of EBH and LBH? In this section, the approach is applied to distinguishing BH from BA clauses. BH and BA are more distinct than EBH and LBH, so it is expected that the LSTM model is able to

8 For an explanation of the concept of "loss" and other technical details related to neural networks, see Appendix E.

produce clearer and more stable results than in sections 6.4.1 and 6.4.2. In the section on variation between EBH and LBH, the prediction accuracy on the validation set was hardly higher than 55%. The research in this section on Hebrew and Aramaic shows that if the languages that one wants to distinguish are more different, the prediction accuracy becomes much higher. This gives more confidence in this way of researching linguistic variation.

If we want to distinguish BH from Aramaic, there are many Aramaic subcorpora from which one can choose. In this research, the most obvious choice is to use the Aramaic portions in the books of Daniel and Ezra, because they have been prepared in a way that is identical to the Hebrew portions of the MT. A disadvantage is that the total number of clauses in these Aramaic portions are limited.[9] A second source of imbalance is that the majority of Aramaic clauses in the MT are Q clauses.[10] With such a limited number of N classes it is difficult to make a fair analysis based on this amount, and that is why only Q clauses are taken into account.

The Hebrew training samples consist of 765 Q clauses from prose texts in books other than Daniel and Ezra. The Aramaic training samples consist of 765 Q clauses from the Aramaic portions of Daniel and Ezra.

The structure of the network is a bit simpler than the network of the previous sections.[11] The model was trained 200 times, each time using a different random sample from the data. This procedure was done for both the phrase level model and the word level model as in the previous sections. The results are shown in figure 6.17 (see next page). The broad black bar in the boxplots show the median values, these are 0.670 for the phrase level and 0.740 for the word level. The accuracy is clearly higher than that of experiments in which Q clauses in EBH and LBH were distinguished, which was generally only slightly higher than 0.5, which can be seen in figure 6.18 (see next page).

Plot 6.19 (see page 173) shows an example of how the loss and accuracy develop during the training process.

---

9 The total number of Aramaic clauses in the MT is 1293.

10 There are 976 Q clauses, 86 N clauses, 1 D clause and 230 ? clauses. As previously done, a Q clause is a clause which has Q as the last character in its value of the feature txt in the ETCBC database. It must be said, that in the ETCBC database the transition from Q to QN is better defined for Hebrew than for Aramaic. What counts most here is that the prediction accuracy increases strongly if the two languages are more distinct.

11 The script with specifications can be found here: https://github.com/MartijnNaaijer/phdthesis/tree/master/Ch6_Sequence_analysis/classify_aramaic_hebrew.

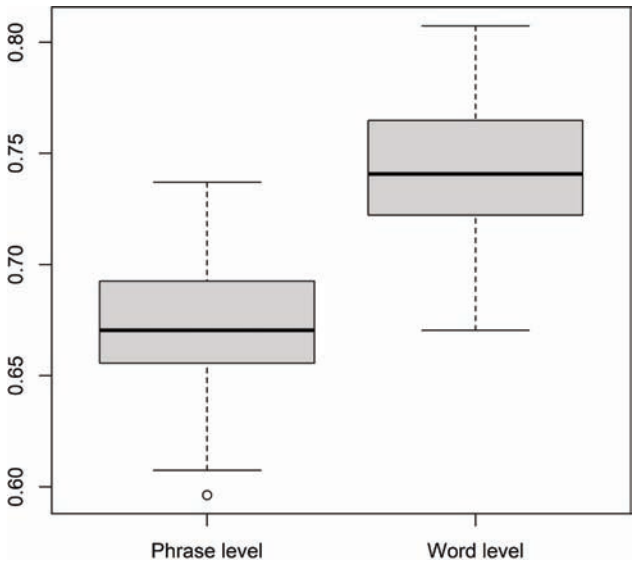

FIGURE 6.17 *Boxplots of predicted accuracies of the classification of Aramaic and Hebrew clauses*

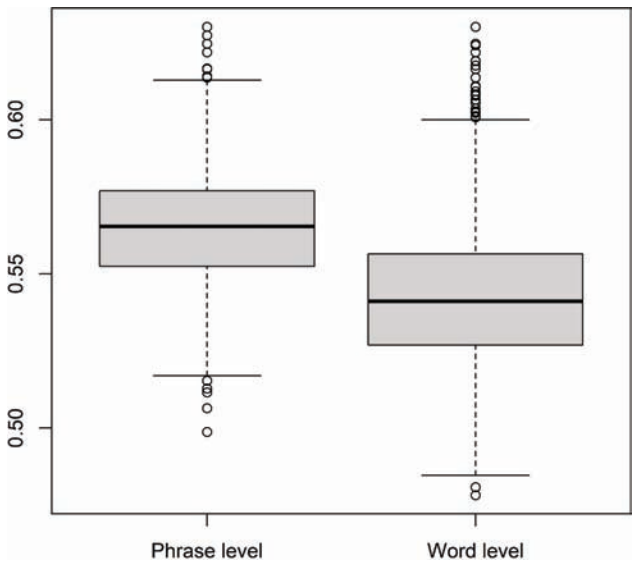

FIGURE 6.18 *Accuracies of the classification of* Q *clauses as* EBH *and* LBH

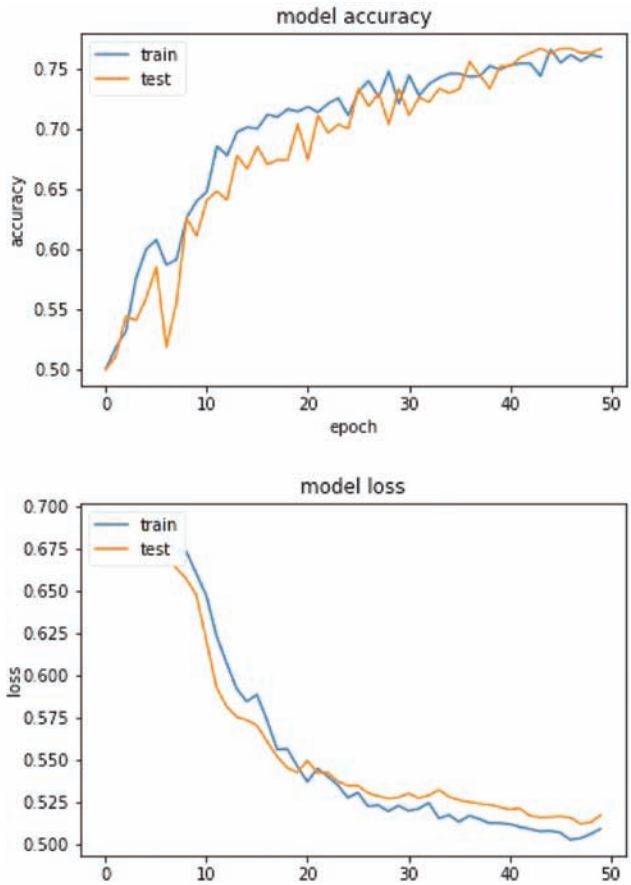

FIGURE 6.19 *Example of loss and accuracy for the phrase level model during training*

The loss decreases gradually towards a minimum, while the accuracy does the opposite, it converges towards a maximum value. Also, the loss of the test set is slightly higher during training than that of the training set, and the accuracy is a bit lower than that of the training set, which is as it should be.

All in all, the experiment in which Aramaic and Hebrew clauses are classified produces good results. Even though Aramaic and Hebrew are closely related languages, the model achieves median accuracies of 67% for the phrase level model and 74% for the word level model. These results are clearly better than the results of the main analysis in sections 6.4.1. and 6.4.2, but they show that the whole approach works well.

6.5. Conclusions

In this chapter, variation between EBH and LBH was investigated using sequence analysis on the levels of phrases and words. The data were split into Q and N clauses and each clause was represented as a sequence of phrase functions in the phrase level analysis or as a sequence of parts of speech in the word level analysis. With such an approach, the language is stripped completely of the content of a story, which makes it easier to focus on syntax instead of on the lexicon. The data were analyzed using an LSTM network. With this network a model was trained, which classifies clauses as EBH or LBH.

In the search for linguistic variation between EBH and LBH, it is important to control for other factors. In this analysis, only prose texts were selected, a distinction was made between quoted speech and narrative, so the only variable of the main variables of this project that was not taken into consideration is main and subordinate clauses. An extra split in the levels of this variable would make the samples small, and the learning process of the model even less stable. For this reason, that split was avoided here.

The analysis shows that it makes sense to distinguish between Q and N clauses in the analysis. There is less variation in Q clauses than in N clauses in relation to the problem at hand.

If the results of the LSTM network of the phrase level analysis are clustered using k-means with k = 2, there is one cluster containing exclusively EBH books and one cluster containing LBH books, but the latter cluster also contains a number of EBH books (Numbers, Deuteronomy, and Kings). Jonah, Ruth and the prose tale of Job are clustered in the former cluster with the EBH books. In the cluster analysis of the word level analysis, Jonah, Ruth, and the prose tale of Job are again clustered with most of the EBH books. Daniel can also be found in this cluster.

So, although most of the EBH books cluster together, and also the LBH books cluster together, there are some books that fall more or less between the two groups, and depending on details, they fall in one cluster or the other. There is some difference between the phrase level and the word level results, but generally, they are similar.

What does this mean for the possible dating of the EBH and LBH books? It is possible that the observed variation between EBH and LBH reflects a diachronic development in BH. From this perspective, Ezra and Nehemiah overall seem to contain the latest language of all books. There is a possibility that in the analysis, the literary relationships between these books caused an apparent increase of the concentration of LBH. Generally, Daniel is considered to be the latest book in the Hebrew Bible, but

this is not visible in the clusters. These results suggest that overall there is a tendency that EBH books contain more EBH language and the LBH books contain more LBH language, but there is no strict distinction, so whereas the traditional distinction between EBH and LBH is visible, it is a distinction with fuzzy boundaries.

What seems to be clear is the status of Jonah, Ruth, and the prose tale of Job. On the basis of the chosen features and approach, there is no other conclusion than that they share most of their characteristics with EBH. What does this mean from the perspective of linguistic dating? If it is accepted that books written in EBH are early, it follows that these are early texts. However, there seem to be various historical, literary, and theological reasons why these texts are exilic or early post-exilic. This can mean two things. First, the texts were written in a time in which the change from EBH to LBH had not taken place yet. Second, in the time that these texts were written, there was more variation in literary Hebrew than is supposed by the chronological model. In the case of these texts, the scholarly literature hints at relationships with various other languages than just EBH and LBH, and it may be the case that these relationships are stronger than those with LBH.

The validation of the approach by distinguishing Hebrew from Aramaic clauses works well. As expected, the test accuracy is higher than in the case of the analysis of EBH and LBH clauses.

Sequence models based on neural networks are a vibrant area of research in the machine learning community, mainly because there are many scientific and commercial applications for computers that understand natural language. The study of BH can profit enormously from the advances made in this field. In this research, sequence classification was used, but there are many other applications, for which sequence modeling can play an interesting role. One example is the encoding of QH texts for the ETCBC database using sequence to sequence (seq2seq) models. Other applications of seq2seq models could be in the study of ancient Bible translations.

CHAPTER 7

General discussion and conclusions

In the Syntactic Variation project and in this thesis, a range of new ideas, tools, and techniques have been explored and applied, to be able to take new steps in the investigation of the problem of syntactic variation in Classical Hebrew.

Open Science

The ideas of Open Science are making their way into Biblical Studies. In scientific research, it is important that the results of a query can be published fully, to make the whole process of scientific research transparent and reproducible. This is true not only for the results, but especially for the data on which the results are based, in this case the Hebrew texts (biblical and non-biblical) and their annotations.

Text-Fabric

The Syntactic Variation project, of which this research is part, is the first project in which the Python package Text-Fabric was used consistently. The Hebrew Bible, as encoded by the ETCBC, was the first dataset in Text-Fabric format. So far, this is the only Open Source text with annotations of the MT available in the field, and the availability of these openly available data is essential for reproducible research and large-scale study of linguistic variation.

The Syntactic Variation project also marks the start of the expansion of the ETCBC database with Hebrew texts other than the books of the MT. During the project, 1QS, Pirqe Avot, Shirata and a number of Hebrew inscriptions were encoded. These texts, together with 1QM, are available in Text-Fabric format in the package "extrabiblical".

Multivariate analysis

In many studies on linguistic variation in BH, individual linguistic observations are immediately related to the grand theories of Biblical Studiesm, like the conventional

linguistic dating approach. Of course, everyone is aware that there are multiple factors influencing concrete linguistic structures, but generally, these have not been taken into consideration together in a systematic way. In the Syntactic Variation project, we have introduced multivariate analysis into Biblical Studies.

Techniques

In this research, I have used a number of new techniques to study linguistic variation in BH. My part in the Syntactic Variation project has been to treat clause structure. A clause can have two types of predication, namely, either one having to do with "being", or one having to do with an activity or state expressed by the verb. These two types behave syntactically differently. In chapter 4, a mixed model was used to study the conditioning of the use of the verb היה in clauses with a subject and predicate complement. Also, in chapter 4, Random Forest and Extreme Gradient Boosting were used to investigate the difference between clauses with and without יש and bipartite and tripartite clauses. From a statistical perspective, chapter 5 is more explorative, treating double object clauses as an example of structures with a verb expressing an activity or state. Finally, in chapter 6, a large-scale investigation was done on the difference in clause structure between EBH and LBH, using an LSTM model. This is a specific kind of neural network, used for the analysis of sequences. With its flexibility, it is able to extract the relevant features automatically from the given sequences.

Inferential statistics vs predictive modeling

The techniques used in this research fall broadly into two categories. These are a statistical approach in section 4.2 on היה clauses and verbless clauses, and predictive modeling in sections 4.3 and 4.4 on clauses with and without יש and on bipartite and tripartite clauses in chapter 6. How do these techniques relate to each other?

For section 4.2, a statistical approach is the most natural solution. Regression analysis with book as random effect is used to find out how the use or non-use of היה is conditioned. For each predictor in the model, the effect size and its significance are obtained, and one also gets information about the explained variance (adjusted R-squared) of the model.

For the research in setions 4.2 and 4.3, at first sight, a statistical approach would

be the best fit. However, because of the low number of clauses in the smallest class, clauses with יש or a tripartite clause, I decided to use a predictive model. This choice has some disadvantages. First, the chosen models do not have a random effect. So, if, for instance, the particle יש occurs with a high frequency in a specific poetic book, but these cases are cluttered in one chapter, this high number is representative neither of the book, nor of poetic books in general. Using a mixed model helps to overcome this problem in a natural way. Using predictive modeling with cross validation helps to solve this problem partly, but not in such a smooth way as with a mixed model. Another disadvantage is that predictive modeling is an indirect way of studying this problem. We want to know how certain linguistic phenomena are conditioned, and this is done by predicting the value of the dependent variable. Variable importance is used to get an impression of the influence of the independent variables, but it remains an indirect approach. On the other hand, predictive modeling has advantages. If a correct prediction is made, the model shows that it "understands" the data. One can see directly that predictive modeling really works.

In chapter 6, predicting the language phase is the most natural solution for the given problem. The question as to which class the language of a text or book, like Jonah, the prose tale of Job, or Ruth belongs, is a classification problem. The same is true for linguistic dating of biblical texts in general: there are two classes, EBH and LBH, and the book of uncertain date needs to be classified to one of these two classes. I have chosen to classify books on the basis of the classification of their clauses, but one could make a different choice.

The dependent variable

If we look at the dependent variable, this research can be split in two groups. In the first group are the sections in which the output variable is some linguistic feature. These are chapter 4, on expressions of "to be", and chapter 5 on verbal valence. The other group consists of chapter 6, in which the dependent variable is the language phase.

The first group, in which the dependent variable consists of linguistic features, is used if one is interested in the conditioning of a particular linguistic feature. In the present research, each of these features occurs frequently (hundreds of times) in the corpus under investigation. This is necessary for the production of meaningful results. However, different linguistic features can be conditioned in different ways. In the present research, for instance, some features seem to occur mainly in quoted

speech instead of narrative (the presence of יש). A feature like the choice of the verb (נתן or שים) in double object constructions does not seem to vary between different discourse types, but there is variation between different books and redactional layers. These results are significant, but if one is interested in how different EBH and LBH are, it is better to look for a method like the one used in chapter 6, in which clauses are classified as being more characteristic of EBH or LBH. With this approach one gets a good impression of how different (or how similar) EBH and LBH are. On the other hand, in chapter 6, little is learned about the structure of individual clauses, which may be dissatisfying from the perspective of someone interested in linguistic features. Overall, both approaches have pros and cons, and the choice for one of these depends on the particular interest of the researcher.

A quantitative approach

What does a quantitative approach add to traditional philological research on linguistic variation in BH? As is clear from previous research on linguistic variation in BH, quantitative arguments often play a role one way or another, but it has hardly been the central focus of research.

In chapter 4, a number of traditional hypotheses about the meaning and structure of clauses was confirmed. A statistical test was used in section 4.2 to show whether or not the observations are based on coincidence. In traditional research, it is much more difficult to make this distinction. Also, the length of clauses and subjects of clauses was shown to play a role in the choice of היה and יש. The latter findings are much harder to study without a systematic quantitative approach for features occurring in large numbers, but the approach used here makes clear what the relationships are.

In chapter 5, the distribution of נתן and שים with double object constructions shows various patterns. There is variation between genres, language phases, and redactional layers, which was revealed by plotting the distribution of clauses. In quantitative research, plotting of results is important, because a clear plot gives a better intuitive understanding of the data than lists of numbers.

The approach used in chapter 6 makes it possible to investigate not only how similar the language of books of uncertain date is to the EBH and LBH books, but also how the EBH and LBH books relate to each other. This is an enormous advance relative to the situation in which only two groups of EBH and LBH books are chosen, and another text or book is dated on the basis of relationships with the LBH books. In the traditional approach of linguistic dating, it occurs frequently, that rare and

idiosyncratic features are selected as being characteristic of LBH. In the analysis of chapter 6, a majority of clauses are selected for the analysis and, of course, these clauses contain rare features, but by studying many clauses, these rare features have a low weight and influence the analysis only to the extent of their relative frequency.

In traditional linguistic dating, the extrabiblical evidence is used to show that a linguistic feature is not only used by a specific author, but that it was used more broadly in later phases of the language. Generally, the extrabiblical evidence is based on the DSS or Rabbinic texts, but sometimes extrabiblical evidence is found in the Targums or other Aramaic literature. If the late feature is found at least once in the extrabiblical literature, the criterion of extrabiblical attestation is satisfied. It remains unclear, however, to what extent the late feature is used thoughout the extrabiblical evidence. If an LBH feature is found in an Aramaic translation of the Hebrew Bible, but it is absent from QH and RH, can one still say that the feature is a sign of diachronic development, or is a different explanation more plausible? In my opinion, in this case it might be better to look for an alternative explanantion of the data. In a situation in which the full data are studied, and visualized with, for instance, mosaic plots, it becomes much clearer whether there is real linguistic development, based on an overview of all the data related to this feature in all the available biblical and extrabiblical texts.

The statistical analysis of the use of היה in clauses with a subject and predicate complement has shown various tendencies in the data. Some of these confirm the traditional idea that this verb adds TAM to the clause. These are, for instance, the increased use of היה in main clauses, in clauses with a mother that is not a verbless clause, and in clauses with a time phrase. היה is used less, if the clause contains a question or interjection phrase, which occur relatively often in clauses in quoted speech. In quoted speech in general, the use of היה does not differ significantly from narrative texts. There is a significantly lower use of היה in poetry than in prose and prophecy, perhaps because the role of time is smaller in poetry.

היה seems to be used to give structure to the clause, because longer clauses tend to contain היה more often. Also, there is an increased use of היה in clauses in which the predicate complement is a PP, which seems to confirm the idea that there is semantic variation in the use of היה, indicating that היה has a broader function than only adding TAM.

The verb היה is used significantly more often in EBH than in LBH, in clauses containing a subject and predicate complement. This lower frequency in LBH is not visible in QH and RH, so LBH has a distinct position here.

The present research confirms the interpretation that the particle יש puts emphasis on the clause. The particle is relatively rare, but in contrast to clauses with a similar structure (clauses with an indefinite subject and a PP predicate complement) without the particle, it occurs predominantly in clauses in quoted speech sections. In the analysis with XGBoost, quoted speech is the most important predictor for the presence of יש. Of the different language phases, the particle occurs most often in RH, which is generally associated with spoken language. Like in clauses with היה, longer clauses tend to contain יש more often than do short clauses.

The research on the tripartite verbless clause suggests a non-copular interpretation of the tripartite clause in BH. An important argument is that the structure is relatively rare (fewer than 200 cases in the MT, which is only a small fraction of the total amount of relevant data), one would expect more cases if it were used as a copula. Another clear sign is that in the analysis using XGBoost, quoted speech is the most important predictor for the use of the tripartite clause, also for a situation in which the fronted subject is resumed and emphasized, quoted speech is the natural environment. The second most important predictor for the tripartite clause is whether the clause is an argument clause. Most of these are object clauses, occurring in quoted speech.

It is theoretically possible that in quoted speech the pronoun functions as a copula, and that the increased use of this copula in quoted speech shows something about the difference between spoken and written Hebrew, but I prefer to use the simplest explanation of the phenomenon. Resumption also occurs in constructions other than tripartite clauses, so with this explanation no new grammatical category needs to be invented or borrowed for BH.

Both clauses with יש and tripartite verbless clauses have a preference for longer subjects. It is possible that יש or the pronoun gives structure to the clause, as was suggested by Driver, in the case of the tripartite clause. On the other hand, especially for clauses with יש, there is a relatively low amount of evidence, so conclusions should be drawn carefully.

In chapter 4, it has become clear that there is a variety of factors influencing the use of היה, יש, and the pronominal copula, and these factors have a varied background. Also, studying these together in one analysis has shown the relative importance of these factors.

In chapter 5, the variation in the distribution between נתן and שים was studied exploratively. The focus was on double object constructions with and without ל, in which the verbs have (more or less) the same meaning, namely: "to make object_1 to be object_2". Double object constructions with נתן and שים occur relatively often in

the Major Prophets: a high frequency with these constructions is to be found in the books of Isaiah, Jeremiah, and Ezekiel. There does not seem to be much variation in the frequency of double object constructions between the main levels of the discourse environment (N and Q), and between main and subordinate clauses.

Within the prophetic books, there is substantial variation of the preference for one of the verbs. In the book of Isaiah, there is a preference for using שים, whereas Ezekiel has a strong preference for using נתן. In the poetic books, the Psalms have an equal use of both verbs, but Job and the Song of Songs prefer שים. Between the prose books, there is substantial variation in the preference for one of the two alternative verbs. Double object constructions with these verbs are nearly absent in the LBH books of Esther, Daniel, Ezra, and Nehemia, but it is common in the book of Chronicles, which has a strong preference for the use of נתן. This is not only reflected in the absolute frequencies of נתן and שים in double object constructions (15 times with נתן and once with שים), but also in parallel passages. In various parallels, Chronicles uses the verb נתן, where the parallel uses a different verb (e.g., שים) or a different construction with the same verb.

In the Pentateuch, on the other hand, most books have a mixed profile. Deuteronomy has a strong preference for using נתן, and the same is true for the P source. J and E, however, have a preference for using שים. The Former Prophets have a mixed profile. Samuel has a preference for שים, but Kings prefers נתן.

A much rarer construction is formed by נתן and שים with double object, in which the second object is introduced by ב. The meaning of the clause is "to make object_1 like object_2". Although it is rare or absent in most books, the distribution of נתן and שים with this construction follows the same patterns as the other double object constructions.

Thus, even though the verbs נתן and שים are synonyms in double object constructions, they are not distributed evenly throughout the Hebrew Bible. Various texts and books have a strong preference for one of the two verbs. The most notable are the assumed sources of the Pentateuch and the book of Chronicles. On the basis of the evidence in the MT, one can conclude that different texts with different backgrounds used a different verb, but it is possible that one verb does not necessarily relate to one specific background. If it is accepted that P is exilic/post-exilic, it is possible that the use of נתן with double object constructions is characteristic of post-exilic Hebrew, because it is used predominantly in Ezekiel, and nearly exclusively in Chronicles.

Linguistic dating is a discipline for which predictive modeling offers a natural solution. A model is trained on features from two subcorpora, EBH and LBH, and predictions are made on unseen data, a text of unknown date, to find out whether

the text's language is more similar to EBH or LBH. In chapter 6, this approach was used to find out what the linguistic relationships are between the separate EBH and LBH books and the relationships between those books and the books of Jonah, Ruth, and the prose tale of Job. The results are compared with the results of traditional linguistic dating.

In this research, an LSTM model was used to classify clauses from Jonah, the prose tale of Job, and Ruth. With the LSTM network, it is possible to model sequence data. Modeling sequences without the need to extract specific features from them is an important step forward in language modeling, because LSTM models are able to detect long term dependencies.

Two analyses were done, one on the phrase level, in which clauses are represented as sequences of phrase functions. The other analysis is done on the word level, in which clauses are represented as sequences of parts of speech. Generally, the traditional distinction between EBH and LBH is visible in this analysis. If the data are grouped in 2 clusters, one cluster contains mainly the EBH books, whereas the other cluster contains most of the LBH books, but there are some EBH and LBH books falling "between the clusters", indicating that there is no sharp distinction between EBH and LBH. This is an important result, because it confirms the traditional distinction between EBH and LBH. On the other hand, it shows that it is difficult to classify the books of undisputed date in the cluster in which they are supposed to belong. Also, results may vary slightly if other features are chosen as input for the model.

In some studies on linguistic dating, Jonah, the prose tale of Job, and Ruth are considered to be written in LBH. This is based on research in which distinct features are selected that are considered to be typical of LBH. This traditional approach has various difficulties attached to it, such as the weight that needs to be given to each of the features. This and various other problems are solved automatically by the LSTM network. Features contribute to the model as far as they contribute to its predictive power. The result of the analysis is that the clause structure of Job, the prose tale of Job, and Ruth is basically that of EBH. Of course, this is based on the way clauses were represented in the analysis, as sequence of phrase functions and as sequence of parts of speech, but this kind of analysis can be extended easily, depending on one's own preferences for features.

How can these results be explained in the light of the traditional linguistic dating approach? In traditional research, a number of features are selected that link a text with the core LBH books. However, as was shown in chapter 2 in the example about Second and Third Isaiah, it is not always clear how features are selected and

to what extent late features are representative of LBH in general. With a large-scale quantitative approach, in which many features are taken into account and weighed, this problem is avoided. Linguistic dating is not only based on the idea that texts share linguistic characteristics, but there is a range of assumptions underlying it. There is no clear empirical evidence that linguistic dating of biblical texts works, which make it a method without any validation. In my opinion, it is more fruitful to analyze linguistic variation from a more descriptive perspective. It is not unlikely that a multidisciplinary approach, in which linguistic, literary and textual history are taken into account may lead to more insight into the linguistic history of Biblical Hebrew.

This research has shown that linguistic variation in BH has a varied background, being partly literary, partly linguistic, and partly historical. There is variation in the use of היה between different genres. Also, double object constructions with the verbs נתן and שים are relatively frequent in prophecy, especially in the Major Prophets, and various books and redactional layers have a preference for one of the verbs with a double object construction.

A substantial part of the variation studied in this research has a linguistic background. יש and the tripartite verbless clause add emphasis, היה adds TAM, the length of a clause and also the length of the subject of a clause influence the use of היה, יש, and the tripartite verbless clause. Also, there is a clear difference in the use of יש and the tripartite verbless clause between narrative and quoted speech. In chapter 6, it was shown that clauses in quoted speech are more homogeneous in variation between EBH and LBH than clauses in narrative.

Finally, part of the background of the variation studied in this research is historical. In chapter 6, it was shown that there is variation between EBH and LBH, although this variation is not so strong that it is possible to draw a clear border between them. Also, it seems that נתן with a double object construction is preferred in late literature.

Recommendations for further research

First and foremost, an important step would be made if more Hebrew texts are included in the research. Encoding texts in the ETCBC format is laborious, and time consuming. Machine learning can likely contribute to encoding texts automatically.

In section 4.2, only clauses with a subject and a predicate complement were studied. This is a substantial number of clauses, but it would be interesting to study

the use of היה more broadly. It was suggested already that the lower use of היה in LBH might be related to a shift in its usage. This could be an increased use of the periphrastic construction, which can be investigated by expanding the dataset with different constructions of clauses with היה, and redoing the analysis with the focus on the variation in its use.

There are various ways in which the research on verbal valence can be extended. On the one hand, extrabiblical texts can shed light on possible diachronic development of these constructions. This research has shown that books can have strongly differing preferences for one of the verbs, and a full study of the the Dead Sea Scrolls and Rabbinic texts may show whether this variation is also visible in post-biblical Hebrew, or that situation in LBH, in which נתן is used nearly exclusively, is extended to post-biblical Hebrew.

A second interesting thing is related to the semantics of the verbs נתן and שים in double object constructions. It is clear that often these verbs have (more or less) the same meaning, but there are also expressions in which only one of the two verbs are used, or in which the second object is (nearly) always introduced by ל. The question arises as to what extent the verbs נתן and שים are interchangeable, and the same it true for the structure of the second object. A more thorough investigation of the semantic content of the clauses with נתן and שים and a double object construction may lead to a better understaning of the semantic similarities and differences between these verbs, and between the difference valence patterns of these verbs.

In this research, the main focus of was on studying נתן as alternative of שים in double object constructions. There are various other constructions that are semantically similar to those studied here. For instance, in 1 Kgs 10:9 the verb שים is used with double object (וישימך למלך, "He has made you king"). Instead of only looking at alternative double object constructions, it could be fruitful to take into account in this case the hiphil of the verb מלך. This would on the one hand complicate the research, because there might be many semantic alternatives. On the other hand, it would make it possible to do some large-scale analysis, because the size of the dataset would increase substantially.

The study of verbal valence is a complex topic, and one could easily fill a whole thesis with a specific group of verbs. One complex group of verbs which has been studied a lot are the verbs of movement. Many variables seem to play a role in the choice of preposition introducing the accompanying locative, and it is worthwile doing such an investigation with more advanced statistical tools than has been done so far, especially because there is a large number of relevant clauses in biblical and extrabiblical Hebew texts.

There are observations in this analysis supporting the diachronic model, but there are also observations that do not fully support it. The question is, then, how to proceed further? Extra-biblical texts may give much more evidence for possible diachronic development of BH in such an analysis. The addition of DSS and Rabbinic texts may show more clearly whether the variation between EBH and LBH is more likely to be diachronic or stylistic, or a combination of these. The use of substantial amounts of extra-biblical texts would require refining the methodology, because a split in just EBH and LBH does not do justice to the nature of the data.

In the present analysis, especially in chapter 6, much linguistic detail was lost by the choice of features (phrase functions and parts of speech). However, there is nothing against including other features, like lexemes. Also, one could include information about the relationships between clauses, such as information on main and subordinate clauses.

There is a rapid development of new machine learning techniques. It is possible to make multi-input LSTM models, in which various sequences are combined as input in one model. In such a model, one could, for instance, make a combined word level and phrase level model. Also, it is possible to use completely different algorithms, for instance, based on Convolutional Neural Networks.

In this research, the focus was on syntactic features, but quantitative research is not restricted to syntax. Recent developments make it possible to study semantics based on the distribution of words and the structure of clauses. Most of the techniques used for this kind of research require a large amount of data, but I expect that new techniques will become available with which it is possible to study semantics based on smaller corpora.

An important factor for the success of new techniques is that members of a research community simply start using them, and adapt them for their own research questions. I have always sensed that the distinction between different areas of research, like the humanities and sciences, is rather artificial. Science is science, and the study of linguistic variation in BH can profit greatly from advances made in recent years in statistics and data science. So far, only a few scholars have tried to integrate these fields, but the infrastructure offered by Text-Fabric makes it much easier. This open source Python package, combined with the analytical capacities of languages like Python and R, can give the field a serious boost.

APPENDIX A

Uncertainty and confidence intervals

In several places in this research, the confidence interval played an important role, and it is crucial for understanding classical statistics. In order to get a better feel for how the confidence interval works, a simulation of random samples is made, to calculate the confidence intervals of the sample mean.

If the value of a certain parameter of interest of a population is estimated on the basis of a sample from that population, the estimation will often not be exactly that of the true value of the parameter of the whole population, because the samples are only partially representative of the population, and there is a certain natural fluctuation between the samples. The question is, how close a certain estimate is to the true value of a parameter. One way to deal with this problem is to calculate the confidence interval around an estimate. Intuitively, it is clear that a large sample from a certain population gives a more reliable estimate than a small one, but how much more reliable is the estimate when the sample grows?

The confidence interval is used to give an indication of the probability of finding the true population value given a sample. A confidence interval is always associated with a certain probability. A 95% confidence interval indicates that if one samples 100 times, the true population value falls within the interval 95 times on average.

Simulating the confidence interval for a normally distributed variable

To illustrate how this works in practice, a simulation is done in which an estimate is made of a parameter of which the true population value is known. From a standard normal distribution (mean = 0 and standard deviation = 1) random samples of size 100 are generated and the confidence interval around the mean of the sample is calculated and this procedure is repeated 100 times. This interval gives an indication of the range in which the true mean can be found. Figure A1 shows the results.[1]

1 The scripts for the simulations in Appendix A can be found here: https://github.com/MartijnNaaijer/phdthesis/blob/master/Appendices/AppendixA.R.

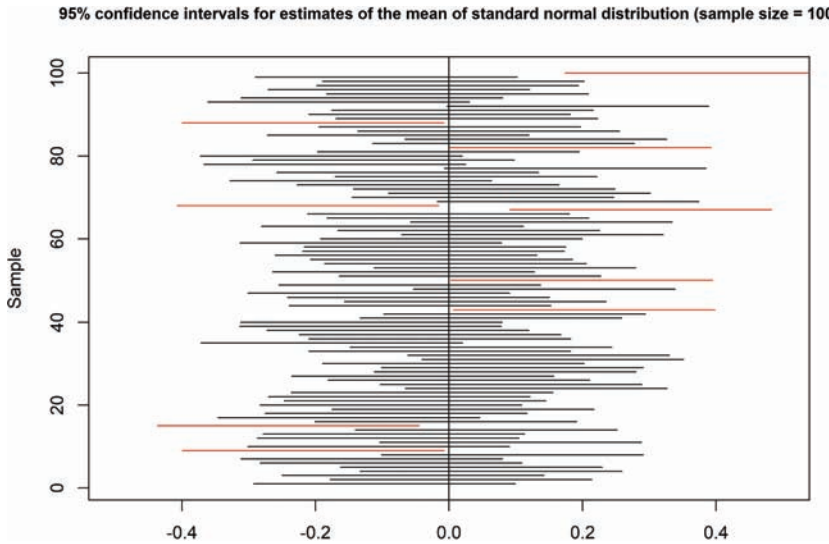

FIGURE A1 *Simulation of 100 95% confidence intervals of the mean of standard normal distribution (n = 100)*

The vertical line indicates the true population mean of 0, and every horizontal line is the confidence interval of one sample. The red lines show those cases in which the true mean does not lie within the confidence interval. In this experiment, there are 6 red lines. The confidence interval which was calculated here is the so called 95% confidence interval.

Of course the size of the 95% confidence interval can be determined experimentally by sampling randomly very often from a population, but if the distribution of the parameter is known, such as in the previous example, it can also be calculated with the following formula:

$$\bar{x} \pm z^*(\sigma/\sqrt{n})$$

Here $\bar{x}$ is the estimated mean of the sample, and $\sigma/\sqrt{n}$ is the so called standard error (se), which is a sample property. σ is the standard deviation of the population (if this is unknown, it can be estimated with s, the standard deviation of the sample), and n is the population size. z is the point on the standard normal distribuation that the probability of observing a value smaller than z is 95%, in the case of a 95% confidence interval. Then $z \sim 1.96$. If the sample size increases, the standard error decreases and the confidence interval becomes narrower. There are similar formula's for the confidence interval of other probability distributions.

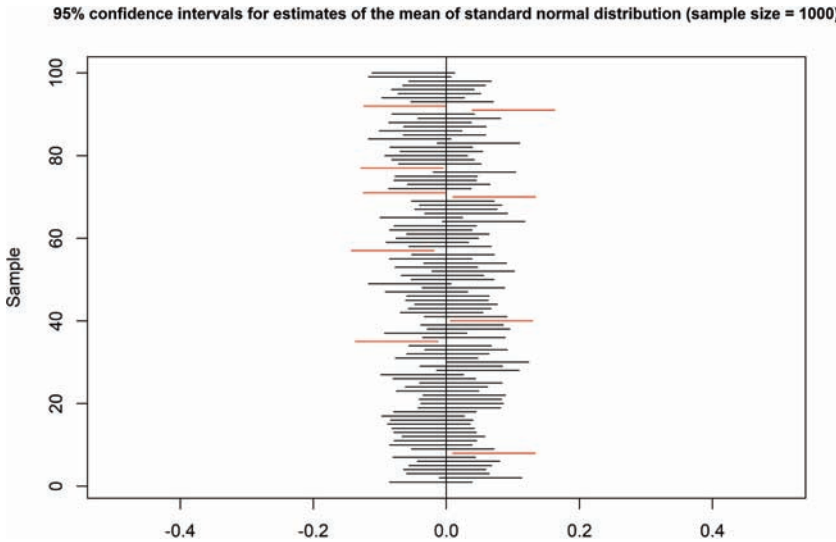

FIGURE A2 *Simulation of 100 95% confidence intervals of the mean of standard normal distribution (n = 1,000)*

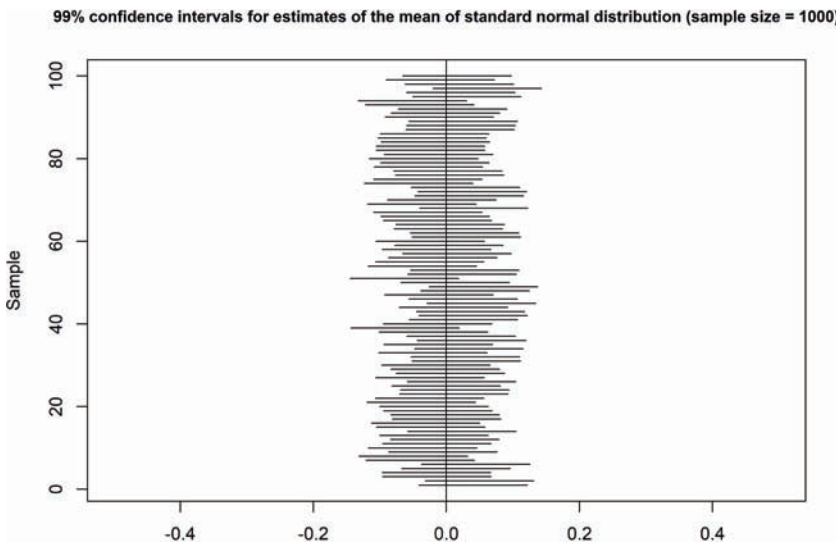

FIGURE A3 *Simulation of 100 99% confidence intervals of the mean of standard normal distribution (n = 1,000)*

So, a more precise estimate of the population value can be made when larger samples are used. Figure A2 shows what happens if samples of 1,000 individual observations are used. There is a drastic decrease of the size of the confidence intervals (they are reduced by a factor of $\sqrt{10}$).

One could think that it is better to take a 99% confidence interval instead of a 95% confidence interval. This means that in 99% of all confidence intervals the true mean is included in the interval. Figure A3 shows the results of this choice. The figure shows indeed that there are more confidence intervals that include the true mean, but it comes at a cost: the intervals are wider, so the individual estimates are less precise.

APPENDIX B

Regression analysis

Regression analysis is a family of statistical techniques, with which the relationships between variables in a dataset are analyzed. In its most basic form, simple linear regression, there is one independent variable (also called input variable or predictor) and one dependent variable (also called output or response variable). In simple linear regression, the variables have continuous, numeric values. If one wants to know the relationship between more continuous predictors and a continuous response variable, multiple linear regression is used. In corpus linguistics, at least in the case of the study of Biblical Hebrew, most variables are not numerical and continuous, but if they are numerical, they are often count variables. This research deals mostly with categorical variables. Categorical variables generally have a limited number of values, and those values do not have a natural order.

If the response variable is a count variable, in general Poisson or Negative Binomial regression is used, and if the output variable has two values logistic regression is used. If a categorical response has more outcomes one can resort to multinomial regression.

Simple linear regression

Simple linear regression is an important basic technique, and various more complicated models are extensions of linear regression. With simple linear regression one analyzes the relationship between one continuous predictor and a continuous outcome variable.

Figure B1 shows a scatterplot of a simulated dataset. Intuitively, it is clear that there is a positive correlation between the variables X and Y.

In linear regression this association is described mathematically as:

$$Y \sim \beta_0 + \beta_1 X + \varepsilon$$

This is the linear model, which assumes that the relationship between X and Y follows a straight line. β_0 is the intercept of the model. This is where the line crosses the y-axis. β_1 is the slope of the model. This is the steepness of the line. ε is the so-called error term. β_0 and β_1 are the parameters that describe the data as good as possible,

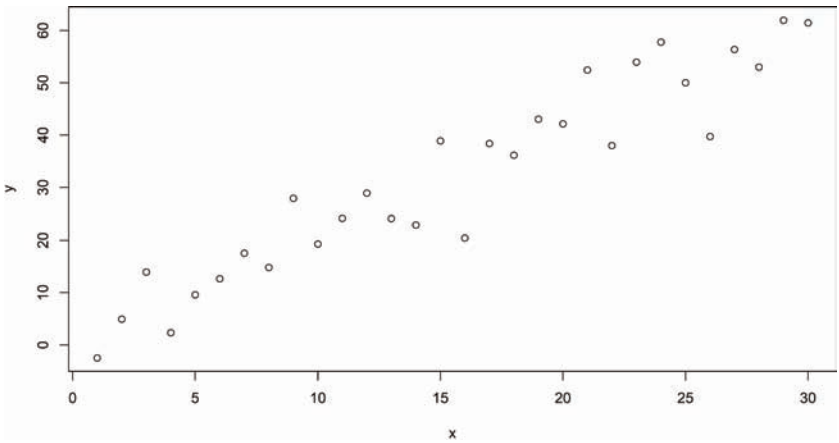

FIGURE B1 *Visualization of simulated dataset with continuous variables* X *and* Y

assuming that the relationship between X and Y is linear. How well the model fits the data is generally measured using ordinary least squares (OLS). The model describes the data with a straight line, but most (or all) of the datapoints do not fall exactly on that line. The distance of an observation to the line is called a residual. See figure B2, which shows the same data as figure B1, but now the linear model is added and the residuals are shown as blue line segments.

The ith residual is the distance of the observation y_i to the value predicted by the model $\hat{y}$. This is $e_i = y_i - \hat{y}$, in which e is the error. The Residual Sum of Squares (RSS) is defined as:

$$RSS = (e_1)^2 + (e_2)^2 + \ldots + (e_n)^2,$$

in which n is the number of observations. The errors are squared, to make all values positive. If this would not be done, the RSS would be 0, because the sum of the negative errors is equal to the sum of the positive errors. Using ordinary least squares the goal is to choose such values of β_0 and β_1, that the RSS is as small as possible. In this case, the model is as close to the data as possible.

There can be a relationship between the variables, but if the noise is strong, the relationship can be invisible. How is it decided the relationship between the variables evaluated in the presence of noisy data? In linear regression, the basic hypothesis is that the model parameters are 0. This is the null-hypothesis. If the null-hypothesis is true, there is no relationship between the data. This null-hyposthesis can be rejected by calculating the so-called p-value. The p-value is the probability of obtaining the

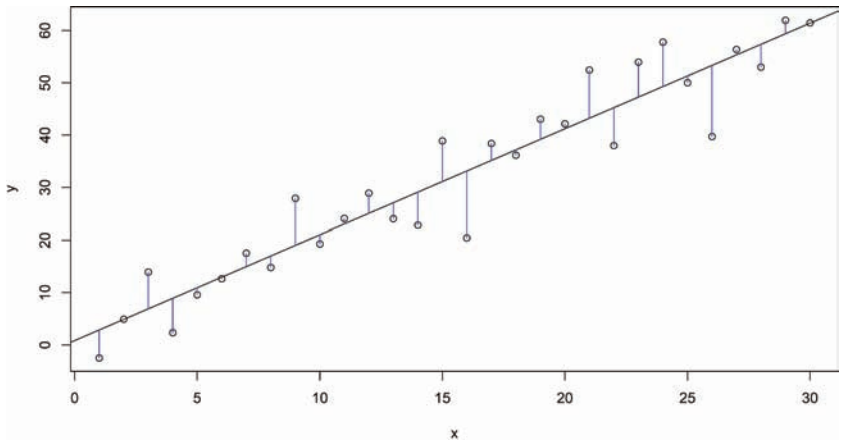

FIGURE B2 *Linear regression model with residuals*

observed results, if the null-hypothesis is true. Using a statistical test one can reject this null-hypothesis if p is smaller than a chosen value. In this research, this value, often called α, is 0.05. So, if this probability p is smaller than 5%, we think that p is small enough to reject the null-hyposthesis. On average, in 95% of the cases this is correct, but in 5% of the cases this is the wrong decision. If p is smaller than 0.05 it is said that there is a statistically significant effect.

It is important to distinguish between statistical significance and practical significance (or relevance). The statistical significance, represented by p, indicates whether or not the effect is based on coincidence or not, but it says nothing about the size of the effect. If the effect is statistically significant, but the effect size is very small, one can say that the effect is not coincidence, but at the same time, it has no practical significance, because there is hardly any variation in the value of the output variable if the value of the significant predictor varies. This (lack of) practical significance is something what a researcher should judge on the basis of his or her knowledge of the field.

The dataset shown in figures B1 and B2, on which these estimates are based have the true values of $\beta_0 = 0$ and $\beta_1 = 2$, and the estimated values of the model are $\beta_0 = 0.85$ ($p = 0.7$) and $\beta_1 = 2.01$ ($p < 0.001$). The p value for the estimate of β_0 is higher than 0.05, which means that the null-hypothesis that β_0 is equal to 0 cannot be rejected. The p-value for the estimate of β_1 is smaller than 0.05, which means that the effect of the predictor X on the output Y is statistically significant. The estimated values of the parameters deviate a bit from the true values. An important question is how close estimated values of parameters are to the true values. For the given simulated dataset,

the true values are known, but in general this is not the case. An indication of the amount of uncertainty of an estimated parameter can be given by calculating the confidence interval of that parameter, as described in Appendix A.

Logistic regression

In the case of a qualitative response variable, linear regression is not the best choice. In the first place, qualitative variables often do not have a natural order as is the case in quantitative variables. Also, if an output variable is qualitative, we want to estimate the proportion (or some transformation of the proportion) of the different outcomes. This proportion should always be between 0 and 1, but in linear regression there is no way to limit predicted values.

This is done with logistic regression. Logistic regression functions similar to linear regression (estimation and significance of parameters), but there is an important difference, which is that the estimated parameters are strictly bounded between 0 and 1. To achieve this, the logit-link function is used. This logit is the natural logarithm of the odds of the estimated parameter.[1]

The logistic model looks as follows:

$$\log(p(x)/1-p(x)) = \beta_0 + \beta_1 x$$

The right side of this formula is identical to the that of the linear model, the left hand side shows the logit transformation. $p(x)$ is the estimated probability of a value of the dependent variable. The value of $p(x)$ can be calculated with the following formula:

$$p(x) = e^{\beta 0 + \beta 1 * x} / \left(1 + e^{\beta 0 + \beta 1 * x}\right)$$

This formula can be used if there is only one continuous independent variable, which is x. Basically, the model described here is still a linear model, but it has a more general use. Therefore, it is called a Generalized Linear Model (GLM).

An example of data that can be modeled with logistic regression is given in figure B3.

1 I do not delve into the details of the logit-link function and its alternatives, like the probit, but more information can be found in James et al. (2009: 131–134).

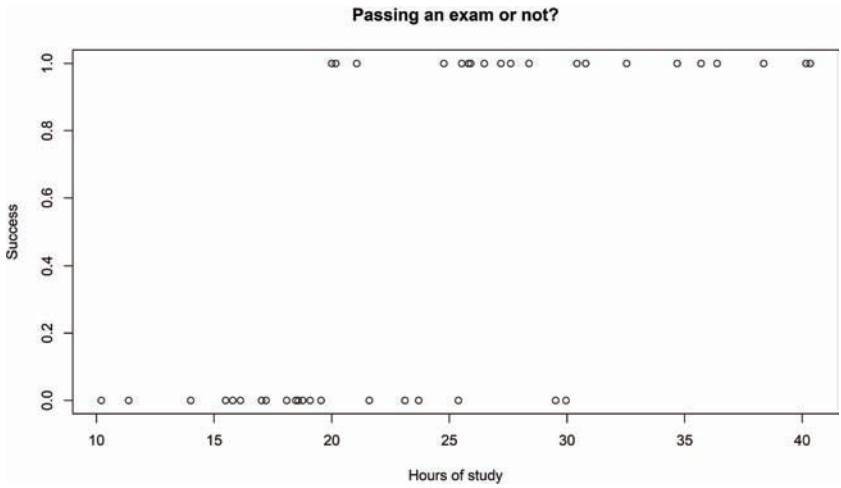

FIGURE B3 *Number of hours of study and the probability of passing an exam*

In this figure, of 40 simulated students, one can see the result of an exam (0 = fail, 1= pass) on the y-axis and the number of hours that a student has studied on the x-axis. The first impression of the plot is that a higher number of hours of study leads to more success. These data can be modeled with a linear model, which is shown in figure B4 (next page, left).

The line in the left figure can be interpreted as the probability of passing the exam. This interpretation creates a problem if the number of hours is lower than 15 or higher than 30, because than the probability is lower than zero or higher than 1, which is not possible. This problem is solved by using a logistic model, which is plotted on the right side of figure B4. Here the distribution of the raw data is shown as a histogram. The red curve never crosses y = 0 or y = 1, so now the curve can be interpreted properly as the probability of passing the exam. The number of hours where the red curve and the horizontal line on probability = 0.5 cross is the number of hours that a student needs to study on average to have a probability of 50% to pass the exam.

The discrete case

In the example above, the independent variable is numeric and continuous. The situation changes if the independent variable is discrete. This situation is relevant for

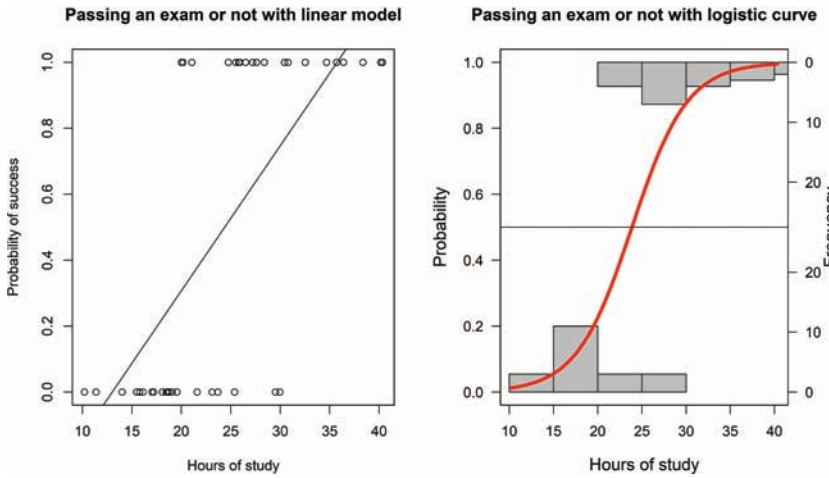

FIGURE B4 *Modeling study success with a linear model (left) and a logistic model (right)*

this research, because the dependent variables are generally categorical. The values of a categorical variable are called its levels, and the purpose of the analysis is to get an estimate of the difference of the probability of the different levels to be able to decide whether or not the null hypothesis, which states that there is no difference between the levels, should be rejected. In a logistic regression model with a categorical independent variable, there is one level defined as the base level, and the other levels are compared with this base level.

Suppose the independent variable X has two values: N and Q. In that case, the model would become:

$$\text{logit}(p(N)) = \beta_0 \qquad \text{if } X = N$$

and

$$\text{logit}(p(Q)) = \beta_0 + \beta_1 \qquad \text{if } X = Q$$

Here, N is the base level, which means that it represents the level of the intercept (β_0), to which the value of β_1 should be added if one wants to know the logit of p(Q). If one works with a categorical variable with three values (suppose these are N, Q and D, again, N is the base level), there would still be an intercept (the base level N) and two extra parameters corresponding to the other two levels. In that case, the one would have:

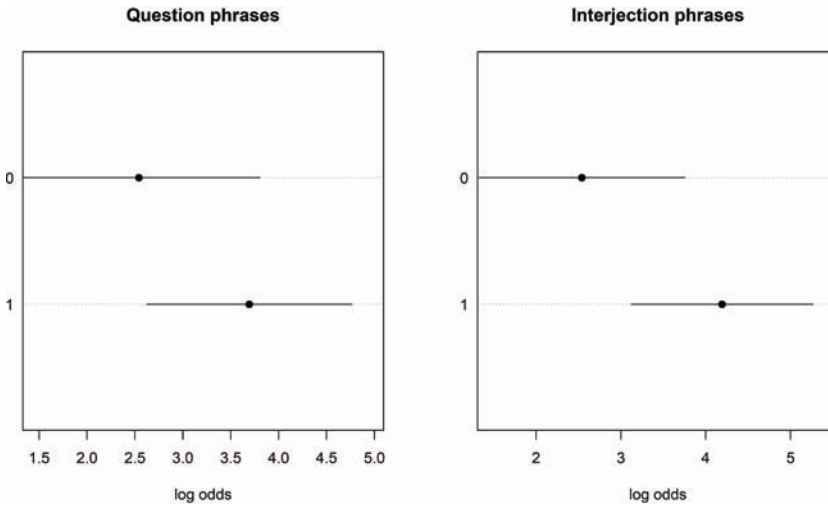

FIGURE B5 *Logistic model of the effect of the presence of certain phrase types on the clause type (verbless clause or* היה *clause)*

$$\text{logit}(p(\text{Q})) = \beta_0 + \beta_1$$
$$\text{logit}(p(\text{D})) = \beta_0 + \beta_2$$

in which β_1 is the parameter for the value Q and β_2 is the parameter for the value D. Note that in both cases, the model describes the difference between the value (Q or D), and the base level (N). An example is given in Figure B5, which is borrowed from section 4.2 in the thesis.

In this figure, the value on the y-axis indicates the absence or presence (respectively 0 or 1) of a phrase with the specified phrase type in a clause. The dots in the figure indicate the logit value of the dependent variable (clause type). The line segments around the dots are the confidence intervals. If the confidence intervals do not overlap, the difference is statistically significant. If they do overlap, it is still possible that the values differ significantly, but that can only be concluded by doing a statistical test.

Mixed models

The logistic model as described above presupposes that the observations are independent of each other. This presupposition is violated in our research, because the textual data are so-called clustered data. In this case, the data are not collected in independent

samples, but they are taken repeatedly from the same environment. Concretely this means, that in the case of clause analysis, two consecutive clauses in a biblical book have a higher probability of sharing characteristics, because they are sampled from the same book, which is the violation of the assumption of independence. This can be avoided by sampling only one clause per book, but this would result in a very small dataset.

A consequence of this clustering of data is that specific characteristics of that individual book are interpreted as a characteristic of some broader group of books. In that case, it is overlooked that the book under consideration is only a single case with its own idiosyncrasies. For example, some linguistic features of Biblical Hebrew of which it is sometimes thought that they are characteristic of LBH, occur only in one or two books.[2] Therefore, it is better to say that these features are characteristic of these books instead of LBH in general. A way to solve this problem is to use mixed models. In a mixed model, one distinguishes between fixed and random effects. Fixed effects are those effects of which the levels exhaust the available levels (for instance, poetry, prose, and prophecy as the main genres in the Hebrew Bible), while in the case of random effects there is only a sample of all the available levels. The most important random effect in this research are the biblical books.[3]

Taking into account fixed and random effects, generally leads to increased p-values in the model, while the effects generally stay the same, if it is compared with an ordinary linear or generalized linear model.

Generalized Additive Models (GAM) and Generalized Additive Mixed Models (GAMM)

An important disadvantage of linear regression is that it is not flexible enough to model non-linear relationships between variables properly. This problem can be solved by using polynomials or other parametric functions. A disadvantage of using predefined parametric functions is that they may lead to rather complicated relationships between the predictors and the target variable.[4] A way to solve this

2 The issue of idiosyncratic features in LBH is discussed in Rezetko and Naaijer (2016 a and b).

3 Instructive examples of mixed effects models in linguistics are given by Baayen (2008: ch. 7), and Levshina (2015: 192–196).

4 Wood (2017: 142–147) shows an example of such a complicated model with polynomials and interactions.

problem is to use a Generalized Additive Model (or GAM). With GAMs it is possible to model non-linear effects in a way that does not need to be defined in advance. A logistic GAM is represented as follows:

$$\log(p(x) / 1\text{-}p(x)) = \alpha + f_1(x_1) + \ldots + f_p(x_p)$$

in which p(x) = probability of Y = 1 given the predictors.

The GAM has more or less the same structure as a GLM. The X's are the independent variables and the f_p's are unspecified smoother functions (Hastie and Tibshirani 2009: 295–296; James et al. 2013: 286–287). Because these smoothing functions are not specified in advance, it is possible to create models with wiggly relationships. Of course, with varying degrees of wigglyness there is the danger of overfitting, but an optimal degree of flexibility between overfitting and underfitting is created by cross-validation. In the formula above, all independent variables are smoothed, but in practice, this is not necessary. One can us smoothers for only one or a few predictors, and it is also possible that the different levels of a categorical predictor each get their own smoother (Hastie and Tibshirani 2009: 297). In R, there are different options to control the smoothing.[5]

An example of a dataset in which a GAM is useful is given in the following figures. Figure B6 shows a simulated dataset based on a third order polynomial with some noise added.

A linear model based on this dataset (figure B7, next page, left) clearly does not fit the data well, so another kind of model has to be looked for. Of course, it is possible to work with a polynomial fit, but if the curve is the random effect of the model, and each individual can have a different kind of curve, than finding the right curve for each individual can be complicated. A better alternative is to model the data with a GAM (figure B7, next page, right). In the figure, the 95% confidence interval is indicated with dashed lines. The model is fitted easily with the code gam(*y* ~ s(x)), in which the function s() indicates that the variable x is smoothed.

In this research, numeric predictors are rare, so in general it is not obvious that smoothers are used. However, the smoothers can also be used to distinguish between fixed and random effects. With this use, the Generalized Additive Model has become

5 In the mgcv package (See Wood 2017 for explanation and examples), the gam function has the smoothing functions s(), te() and ti(). Further parameters in these functions are k for the number of knots, d is for specifying the scale of interacting variables and bs is for specyfing the type of underlying base functions. For s() the default is thin plate regression spline, for te() and ti() the default is cubic regression spline.

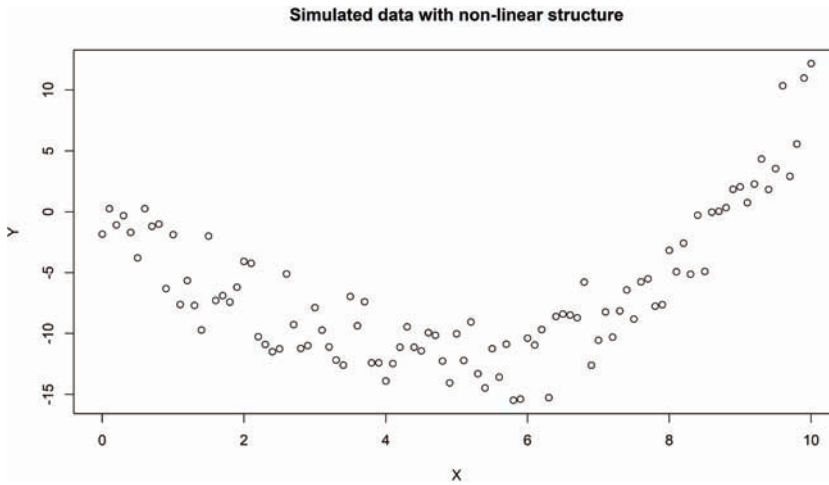

FIGURE B6 *Simulated data with a non-linear structure*

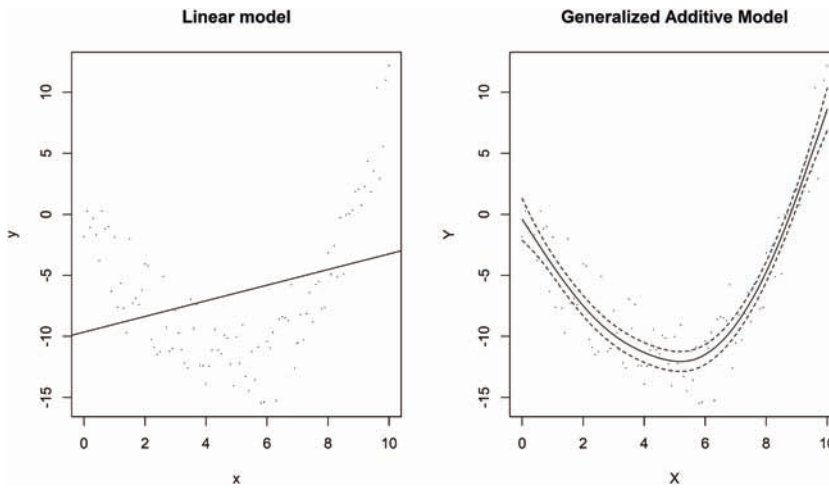

FIGURE B7 *Data with a non-linear relationship modeled with a linear model and a* GAM

a Generalized Additive Mixed Model (GAMM). The data in this research are clustered in books, which violates the independence assumption of the GLM. The data are not just clustered within books, they are also found in a time series, if a biblical book is considered as a time series of clauses. This may cause autocorrelation of the residuals. In this research, a random smoother is used over the variable time for each separate book under investigation. This random smoother includes a random slope and random intercept and in the models the trend of the outcome variable throughout the books. A further advantage of the use of the GAMM is often a reduced autocorrelation of the residuals (Baayen et al. 2016).

APPENDIX C

Cross validation and the ROC curve

An often-recurring problem in machine learning is overfitting. Overfitting occurs if a model fits a dataset on which it was trained very well, but unknowingly, it models not only the underlying patterns in the data, but also its idiosyncrasies. The result may be that if predictions are made on new data, the accuracy is much lower than in the case of predictions on the training dataset. In this case the model does not generalize well to unseen data. A good model does not only fit the data well on which it was trained, but it can also be generalized to other, unseen data. In order to avoid "modeling" these idiosyncrasies, it is necessary to see how well the model fits unseen or 'new' data. This can be done by validation of the model, which is done by splitting the data in two parts before training in a training set and a test set. In general, the training set consists of about 80% of the data, but a different choice can be made. The model is trained on the training set and with the model predictions are made on both the training set and the test set. Ideally, the accuracy of the predictions on both sets is more or less equal, but if the accuracy on the train set is much higher, there is overfitting and one should improve the model somehow.

In this research, k-fold cross validation is used. The k generally stands for five or ten, but in principle it can stand for any number larger than one. In the case of five-fold cross validation, the dataset is split in five parts. The model is trained on four parts, after which the model is tested on the fifth part. This procedure is repeated five times, each time with a different part of the data as test set. With this approach, all the observations are used in a test set in one of the five folds. Figure C1 shows five-fold cross validation.

A special case of k-fold cross validation is Leave One Out Cross Validation (LOOCV). In LOOCV there are as many folds as there are observations in the dataset. In every fold, one observation is used as test dataset. An advantage of LOOCV is that in every fold nearly all the data can be used for training, a disadvantage is that if the number of observations is large, training all the models can take a long time.

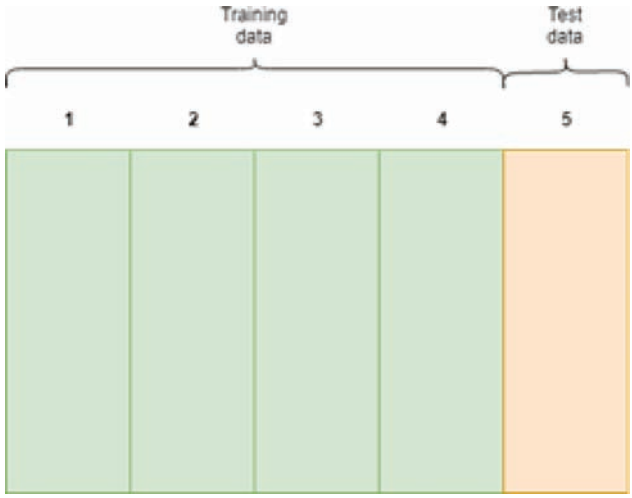

FIGURE C1 *Five-fold Cross Validation*

The ROC curve

The Receiver Operating Characteristic (ROC) curve is a way to evaluate the performance of a predictive model visually. It is used for datasets with a binary output variable. One value is defined as positive, the other is negative. The goal of a predictive model is to predict with a high accuracy on unseen data, but this can be evaluated in various ways. Often, the positive value occurs in a low frequency relative to the negative value, as is often the case in diagnosing diseases. In this case, it can be important to focus on the cases that are predicted as positive. The True Positive Rate (TPR) is the fraction of the positives in the test set that are classified correctly. The TPR is also called the sensitivity. The False Positive Rate is defined as 1 minus the specificity. The specificity is the fraction of negatives that are classified as negative. Figure C2 shows an example of a ROC curve. It consists of a plot in which the TPR is plotted on the y-axis against the FPR on the x-axis, for various values of the cut-off value.

A model which classifies all cases in the test set perfectly has a curve which moves from the lower left corner via the upper left corner to the upper right corner. A model which classifies randomly goes from the lower left corner to the upper right corner on the diagonal line in the figure. Such a model has no predictive power. This means that one wants to train a model in such a way, that the ROC curve is as high as possible in the figure.

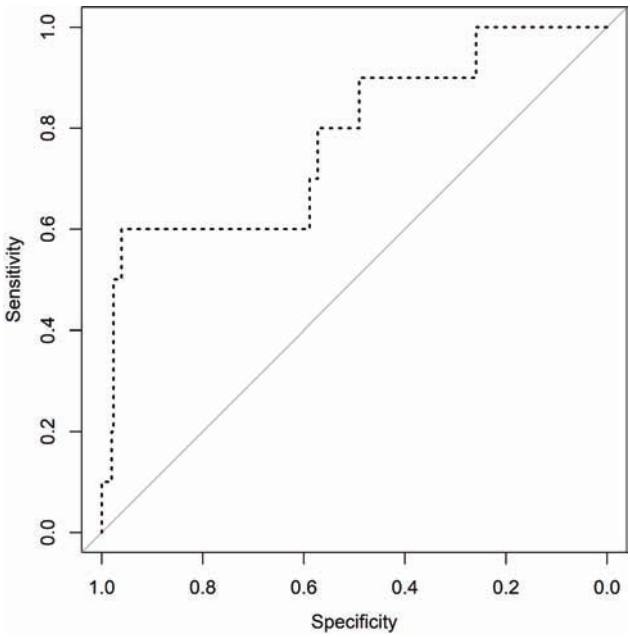

FIGURE C2 *Example of a* ROC *curve*

APPENDIX D

Tree-based models

Decision Trees

Random Forest is a technique for regression and classification based on a decision tree. A decision tree is a supervised learning technique in which a dataset is split in subsets iteratively. These splits take place in such a way that the subsets are as homogeneous as possible in relation to the dependent variable. For example, one can make a spam detector (a favorite example in tutorials). There is a labeled dataset containing 1,000 emails, of which 500 are spam and 500 are ham. Whether an email is spam or ham is the dependent variable. There are several characteristics of an email that make it more likely to be spam. Three of these characteristics (the independent variables) are the presence of suspicious words, an unknown sender or the presence of images in the email. All the emails are classified according to these three variables and the result is as follows:

TABLE D1 *A spam/ham dataset*

	Suspicious words		Unknown sender		Contains images	
	Yes	No	Yes	No	Yes	No
Spam	500	0	400	100	250	250
Ham	0	500	200	300	250	250

In this dataset, it is clear that the best predictor of spam is the presence of suspicious words in an email, because if the email contains suspicious words, one can be sure that the email is spam. If an email contains images, there is a 50% chance that it is spam and 50% chance that it is ham, so this predictor does not provide any useful information about the dependent variable. If the sender is unknown there is a chance of 2/3 that the email is spam (400 / 400+200), and if the sender is known, there is only a chance of 0.25 that the email is spam, so this is also a relatively good predictor.[1]

1 For information-based learning and tree models, see also Kelleher et al. (2015), chapter 4.

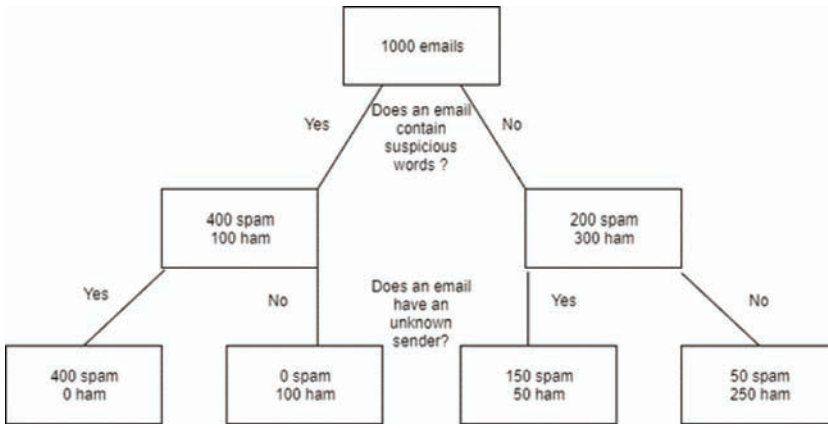

FIGURE D1 *A tree-based split of spam and ham*

In this simple example, it is clear that the best question one can ask, if one wants to know whether an email is spam is: Does the email contain suspicious words? Often, the criterion with the best predictive value does not split the dataset perfectly, so it may be useful to ask more questions of the dataset, starting with the question that gives the best split first, then following with the question that gives the next best split and so on. In the case of the spam/ham dataset, this could look as follows. Suppose that the variable suspicious words does not give a perfect split, we could get the following structure, see figure D1 on the next page. The figure shows that a nearly perfect split is reached after two consecutive questions. This is exactly what happens in the decision tree.

Mathematical background

Intuitively, it is immediately clear that the criterion of suspicious words works better to split the dataset in spam and ham, but how does this work mathematically? In information theory, the notion of cross-entropy is used to describe the impurity of elements in a set. Cross-entropy is defined as:

$$H[p] = -\sum_{i=1}^{k} p_i \log p_i$$

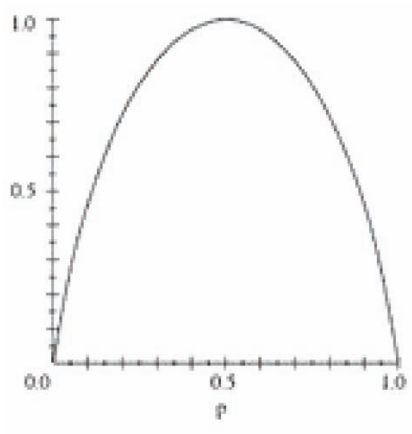

FIGURE D2 *Binary cross-entropy*

In this formula, H is the cross-entropy, which is the information of a system, k is the number of different classes in the dataset (in the case of the spam dataset k = 2, because it can be spam or ham) and p_i is the probability of class i. In the case of a binary variable, this looks as follows, see figure D2.

If the probability of both alternatives is 0.5 the cross-entropy has its maximum, its value is one, which can be calculated easily:

$$-(0.5*\log_2 0.5 + 0.5*\log_2 0.5) = 1$$

If the probability of one of the alternatives in a binary system is 0, in which case it does not occur, the entropy is 0. The system is completely homogeneous then, because only one of the alternatives is present. This can be calculated with:

$$-(0*\log_2 0 + 1*\log_2 1) = 0$$

In a pure or homogeneous system, the entropy has its minimal value, which is 0, and it is higher in the case if the system is impure, with a maximum at 0.5. The decision tree is based on this idea, because if a dataset is split on the basis of a certain criterion, it judges the split on the basis of the information gain obtained with the split. The information gain is the entropy difference between the situation before and after the data split.

In the example of the situation in table D1, the entropy before the split in two groups was 1, because the probability that a randomly chosen email is spam is 0.5.

After the split in two groups in three different ways the situation is as follows. After the split on the basis of the criterion of suspicious words the entropy is:

$$0.5(-(0*\log_2 0 + 1*\log_2 1)) + 0.5(-(1*\log_2 1 + 0*\log_2 0)) = 0$$

After the split on the basis of the criterion of an unknown sender the entropy is:

$$0.6*(-(1/3*\log 21/3 + 2/3*\log 22/3)) + 0.4*(-(3/4\log 23/4 + 1/4\log 21/4)) = 0.875$$

Finally, after a split on the basis of the presence of pictures the entropy is:

$$0.5(-(0.5*\log_2 0.5 + 0.5*\log_2 0.5)) + 0.5(-(0.5*\log_2 0.5 + 0.5*\log_2 0.5)) = 1$$

Before each split the entropy was 1, and the lowest entropy is achieved by splitting the data on the basis of the presence of suspicious words. In this case, the entropy difference, or the information gain, is 1–0 = 1. Note that after the split based on the presence of pictures, there is no information gain at all. The highest information gain is the criterion on the basis of which a decision tree splits the data efficiently (James et al. 2013: 311–314).

Random Forest

Random Forest is based on decision trees. As is often the case in machine learning, the decision tree can easily lead to overfitting. A way to solve this is by using Random Forest. Instead of making one decision tree, with Random Forest a large number of trees is created. Each tree is based on a random subsample of the features in the dataset, and the results of all the trees are averaged. Random Forest can be seen as a kind of "wisdom of the crowd", because the model is based on the votes of many, slightly different, single trees. This kind of model based on a collection of models is called ensemble models. A single decision tree can be visualized easily in the way of the visualization in figure D1. A Random Forest model is harder to visualize, because it is based on multiple decision trees (James et al. 2013: 319–321).

Extreme Gradient Boosting

Another extension of the decision tree is Extreme Gradient Boosting (XGBoost). It is one of the various boosting algorithms. Just like Random Forest it is based on tree learning. There is also a fundamental difference. In a Random Forest, the trees are independent of each other, but in the case of XGBoost, trees are created in a sequence, whereby a tree depends on the tree made in the previous iteration. In every new tree, it is tried to improve classification by giving misclassifications in the previous tree a heavier weight. This approach leads to a sequential improvement of classification (James et al. 2013: 321–323).

APPENDIX E

Neural Networks

Introduction

An artificial neural network (ANN or simply NN) is a mathematical structure, which can be used for numerous predictive tasks. Neural networks have been the subject of intensive research in the past decennium, and have become one of the most used types of algorithms in computer vision and natural language processing.

The simple linear regression model that was considered in Appendix B has two parameters, the intercept and the slope of the straight line through a cloud of points. In contrast, a neural network can easily have hundreds of thousands or even millions of parameters. This gives it the ability to model strongly nonlinear phenomena like pictures or text. This complexity of the neural network has some downsides. Complex models easily overfit, so validation with a test set is important. Also, models with many parameters are difficult to interpret. In general, it is difficult to see how the features of the data are related to the parameters, so neural networks are often seen as "black-box models". There is an input, the data, and there is an output, which is a prediction, but what happens in between, and how that should be interpreted, remains unclear. Making more interpretive neural networks is the subject of research, so it is expected that progress will be made on this in the near future. At a high level the following happens in a neural network, see figure E1.

On the left side of the figure the training samples enter the network, they are processed in the network and a prediction is made on the basis of the data and the parameters of the network. This step is called forward propagation. Then all the predictions are compared with the true values of the output variable of the samples in the training set. The difference between the predictions and the true values is called the loss. Based on this loss the parameters of the model are updated. This is done with a technique called gradient descent. The update of the model makes the next prediction a bit better. The update step is called backpropagation. One iteration of forward and backward propagation together is called an epoch. The network is trained in a sequence of epochs, until training is stopped, or until no improvement of the prediction can be observed anymore.

Neural networks consist of layers of neurons. In figure E1, there are three layers, an input layer, a hidden layer, and an output layer. A network can have an arbitrary

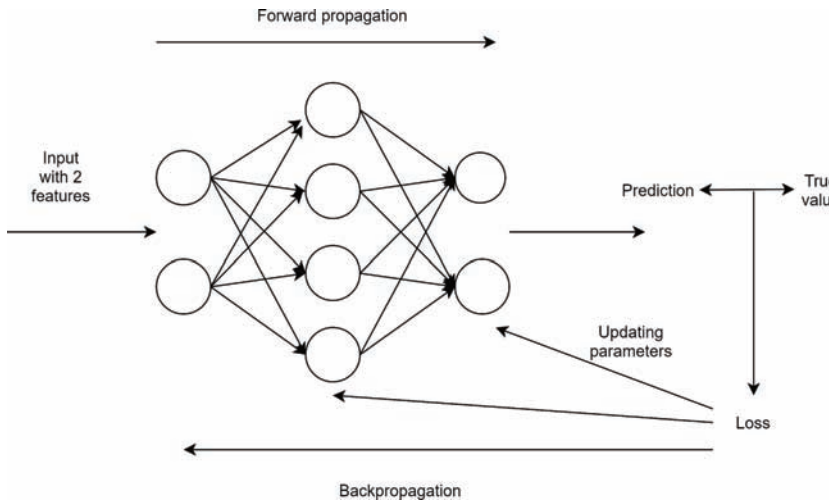

FIGURE E1 *Structure of a neural network*

number of hidden layers, and if it has many hidden layers, it is called a deep neural network. The network in figure E1 is called fully connected, because every neuron in a layer is connected with every neuron in the next layer.

The data flow through the neurons of the network and at the right side a prediction is made, which consists of two numeric values in the figure. In general, in the case of a categorical output variable the output layer has as many output neurons as there are values in that variable. Every output neuron represents one of the values of the output and the numbers that are the result of the calculation are the predicted probabilities of the values of the output variable.

In each neuron in the network, an elementary calculation is performed. Figure E2 shows an example of what happens in a single neuron.

In this neuron, the input value x has the value 4. It is multiplied with the parameter w = 0.5. Every connection in the network is associated with a single parameter. In neural networks, parameters are generally called weights, hence the letter w. Then to the result another parameter, called bias or b, is added. To the result, 3 in this case, a non-linear function (f) is applied. There is a variety of non-linear functions that can be used here, such as the sigmoid and the hyperbolic tangent. In the figure, the ReLU is used. ReLU (Rectified Linear Unit) is a very simple function. If the input of the ReLU is lower than 0 the output is 0, else the output is the number itself, so in this case the output is 3. This type of neuron is also called a perceptron, and the network consisting of these neurons is called the MultiLayer Perceptron (MLP)

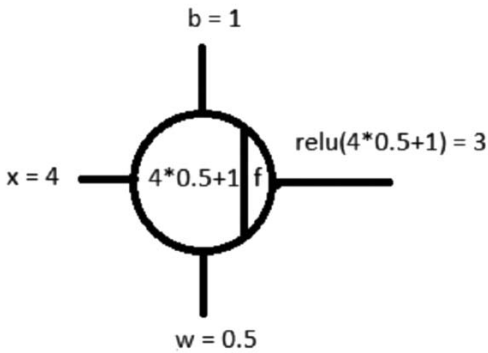

FIGURE E2 *A single neuron*

The weights of the model are generally initialized randomly, which means that the first predictions are also random, but in the process of consecutive epochs the predicted values come closer and closer to the true values.

Gradient descent

As indicated, neural networks learn using a technique called gradient descent. The purpose of gradient descent is to find the minimum of the loss function in an iterative process. This is done by calculating the partial gradient of the loss with respect to the model weights. Then the gradient is multiplied by the learning rate. The learning rate is a hyperparameter of the model, which can be chosen by the researcher. Figure E3 shows the effect of the choice of the learning rate on the optimization of the model.

The loss function has a cone shape, and an optimized model has a loss that is as low as possible. After the first epoch, the overall loss is reduced a bit in the direction of the minimum loss. If the learning rate is too big, the minimum will be overshot. On the other hand, a learning rate that is too small will lead to a slow learning process, which take too long in practical situations. A good learning rate reaches the minimum loss in a number of epochs that is not too high. How many epochs this is depends on the structure of the data and the algorithm, but often a learning rate of 0.001 is chosen.

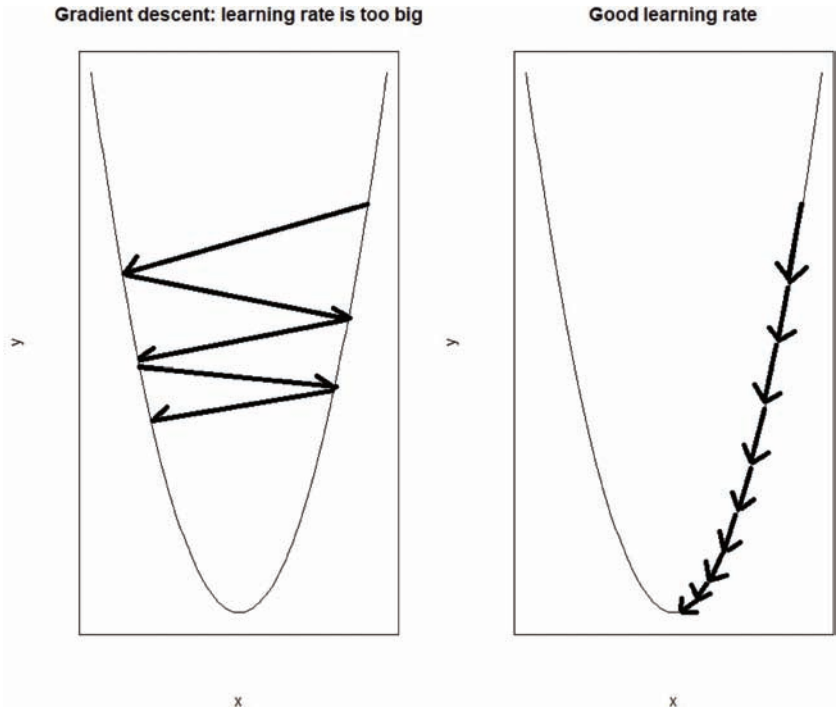

FIGURE E3 *Gradient descent and the choice of the learning rate*

The Recurrent Neural Network

The Recurrent Neural Network (RNN) is a type of network, in which data are modeled that occur in a sequence. A sequence can be a numeric time series or speech, but also text. In the fully connected neural network, it is assumed that all data samples are independent of each other. The RNN cell (in an RNN the neurons are often called cells), however, considers that the input of the present time depends on what has happened previously, because the output of a cell at time t serves as the input of the same cell at time t+1, together with a new observation. Figure E4 shows what happens in a recurrent cell.[1] Left and right of the green arrow the same unit is drawn, but on the right-hand side the cell is unrolled in time. The input is represented by x and o is the output. U, V, and W are the weights of the model and s is the hidden state or

1 The picture was copied from http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/ on 25 January 2018.

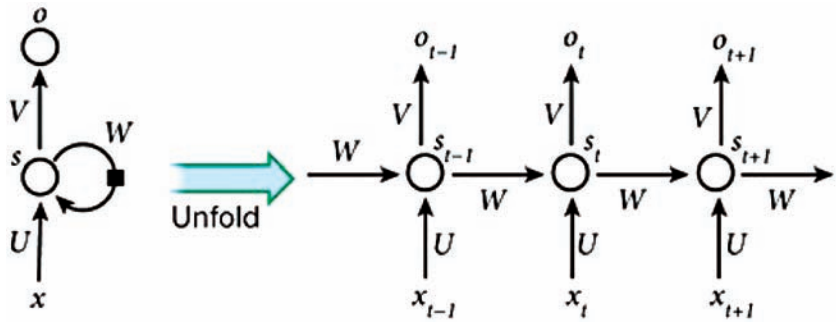

FIGURE E4 *One cell in the Recurrent Neural Network*

the memory of the model. On the right-hand side, in the unrolled representation of the recurrent cell one can see three different time steps. On each of these three different time steps, the input can for instance be a word. In that case, three words can be consecutive words in a clause. On every time step, the cell has two inputs. The first input, x, represents the new word in a clause. In that case, the figure shows the processing of three subsequent words on times t-1, t and t+1. The other input of the cell is W, which is the weight produced on the previous time step. This weight is the element which contains memory of previous time steps and gives the RNN its strength in analyzing sequential data.

It should be noted that the neural network is only capable of processing numbers, and not things like words. Therefore, if the data under consideration are non-numeric, they should be converted to numbers first.

The LSTM Network

A long standing problem in the development of the RNN has been how to train its weights properly. RNN's often suffer from the vanishing or exploding gradient problem. The gradient is the steepness of the learning curve as shown in figure E3. In the case of a vanishing gradient, the gradient is so small that the update of the weights hardly changes them and the learning task is not accomplished. In the case of an exploding gradient, the opposite is the case, the gradient is very high and convergence will not be achieved. Solutions for these problems are offered by the GRU (the Gated Recurrent Unit) and the LSTM (Long Short-Term Memory) networks. Characteristic of the LSTM cell are its gates. It has an input gate, output gate, and forget gate. These gates are control the information in the cell, organizing which information should be remembered, and which information can be forgotten.

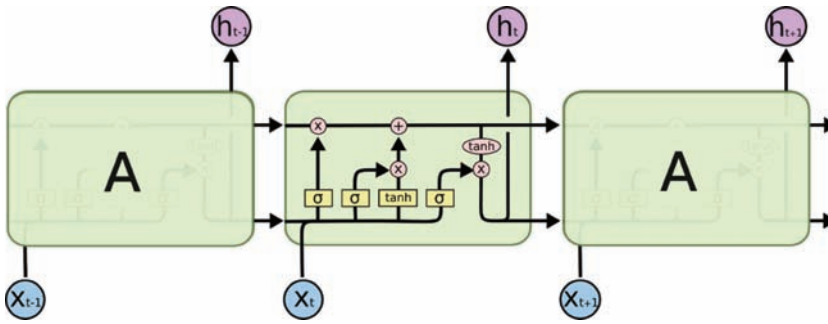

FIGURE E5 *An LSTM-cell*

Figure E5 shows an LSTM-cell on three time steps.[2] The basic sequential structure as in the ordinary RNN is visible, but the cell itself has a more complex structure

The model in chapter 6 of this research is a so-called many to one model, more specifically a sequence classification model. The input consists of a clause, which is represented as a sequence of words or phrases, and the output is a class, which is simply whether the clause is classified as EBH or LBH.

The LSTM network in this research consists of four layers. The first layer is an embedding layer, which is followed by 2 LSTM layers and a final dense layer with one neuron. The embedding layer is used often in neural networks for text analysis. An embedding layer can be used alone to learn word embeddings, but in can also be used as the first layer of a deeper network, as in this analysis. The activation function used in both LSTM layers is the ReLU (Rectified Linear Unit). The final layer consists of only one cell with sigmoid activation, as is usual in classification tasks with only two classes.

Next to the sequence classification task in this research, there is a whole range of interesting applications in which LSTM cells play a role. Examples are image captioning and machine translation. Similar to machine translation is POS tagging, which also uses sequences as both input as output. For BH, this is done in a blog post I have written for the ETCBC website.[3]

2 The figure was copied from Colah's blog, where the structure of the LSTM network is explained well: http://colah.github.io/posts/2015-08-Understanding-LSTMs/?fbclid=IwAR20MVF0HjzRe4fIkd_BCSE31h952JdtS7c0j1SPcnsagdsxg4KDOwabov4.

3 http://etcbc.nl/computational-linguistics/new-text-fabric-module-the-dead-sea-scrolls.

APPENDIX F

The verbs נתן and שים with double object constructions in Genesis-Numbers

The following clauses contain the verb נתן with a double object construction (direct object + second object with or without ל). The letter after the verse (J, E, P) indicates the source according to Driver (1892b).

Gen 16:3 P וַתִּתֵּן אֹתָהּ לְאַבְרָם אִישָׁהּ לוֹ לְאִשָּׁה, And gave her to her husband Abram as a wife.

Gen 17:5 P כִּי אַב־הֲמוֹן גּוֹיִם נְתַתִּיךָ, For I have made you the ancestor of a multitude of nations.

Gen 17:6 P וּנְתַתִּיךָ לְגוֹיִם, And I will make nations of you.

Gen 17:8 P וְנָתַתִּי לְךָ וּלְזַרְעֲךָ אַחֲרֶיךָ אֵת | אֶרֶץ מְגֻרֶיךָ אֵת כָּל־אֶרֶץ כְּנַעַן לַאֲחֻזַּת עוֹלָם, And I will make for you, and for your offspring after you, the land where you are now an alien, all the land of Canaan, for a perpetual holding.

Gen 17:20 P וּנְתַתִּיו לְגוֹי גָּדוֹל, And I will make him a great nation.

Gen 27:37 J וְאֶת־כָּל־אֶחָיו נָתַתִּי לוֹ לַעֲבָדִים, And I have made for him all his brothers as servants.

Gen 29:24 J וַיִּתֵּן לָבָן לָהּ אֶת־זִלְפָּה שִׁפְחָתוֹ לְלֵאָה בִתּוֹ שִׁפְחָה, Laban made his maid Zilpah for his daughter Leah as her maid.

Gen 29:28 E וַיִּתֵּן־לוֹ אֶת־רָחֵל בִּתּוֹ לוֹ לְאִשָּׁה, Then Laban made for him his daughter Rachel a wife.

Gen 29:29 P וַיִּתֵּן לָבָן לְרָחֵל בִּתּוֹ אֶת־בִּלְהָה שִׁפְחָתוֹ לָהּ לְשִׁפְחָה, Laban made his maid Bilhah for his daughter Rachel as her maid.

Gen 30:9 E וַתִּתֵּן אֹתָהּ לְיַעֲקֹב לְאִשָּׁה, And she made her as a wife for Jacob.

Gen 48:4 P וְנָתַתִּי אֶת־הָאָרֶץ הַזֹּאת לְזַרְעֲךָ אַחֲרֶיךָ אֲחֻזַּת עוֹלָם, And I will make this land for your offspring after you as a perpetual holding.

Exod 6:8 P וְנָתַתִּי אֹתָהּ לָכֶם מוֹרָשָׁה, I will give it to you for a possession.

Exod 7:1 P נְתַתִּיךָ אֱלֹהִים לְפַרְעֹה, I have made you like God to Pharaoh.

Exod 16:15 P אֲשֶׁר נָתַן יְהוָה לָכֶם לְאָכְלָה, Which YHWH has made for you as food.

Exod 18:25 E וַיִּתֵּן אֹתָם רָאשִׁים עַל־הָעָם שָׂרֵי אֲלָפִים שָׂרֵי מֵאוֹת שָׂרֵי חֲמִשִּׁים וְשָׂרֵי עֲשָׂרֹת, And he appointed them as heads over the people, as officers over thousands, hundreds, fifties, and tens.

Exod 23:27 J וְנָתַתִּי אֶת־כָּל־אֹיְבֶיךָ אֵלֶיךָ עֹרֶף, And I will make all your enemies turn their backs to you (ערף, "neck" is direct object).

Lev 6:10 P חֻקָּם נָתַתִּי אֹתָהּ מֵאִשָּׁי, I have made it their part of my fire offerings.

Lev 7:32 P וְאֶת שׁוֹק הַיָּמִין תִּתְּנוּ תְרוּמָה לַכֹּהֵן מִזִּבְחֵי שַׁלְמֵיכֶם, And the right thigh from your sacrifices you make an offering for the priest as an offering of well-being.

Lev 7:34 P וָאֶתֵּן אֹתָם לְאַהֲרֹן הַכֹּהֵן וּלְבָנָיו לְחָק־עוֹלָם מֵאֵת בְּנֵי יִשְׂרָאֵל, And I have made them for Aaron the priest and for his sons as a perpetual due from the people of Israel.

Lev 14:34 P אֲשֶׁר אֲנִי נֹתֵן לָכֶם לַאֲחֻזָּה, Which I make for you a possession.

Lev 26:31 P וְנָתַתִּי אֶת־עָרֵיכֶם חָרְבָּה, And I make your cities a ruin.

Num 5:21 P יִתֵּן יְהוָה אוֹתָךְ לְאָלָה וְלִשְׁבֻעָה בְּתוֹךְ עַמֵּךְ, The Lord make you an execration and an oath among your people.

Num 8:19 P וְאֶתְּנָה אֶת־הַלְוִיִּם נְתֻנִים | לְאַהֲרֹן וּלְבָנָיו מִתּוֹךְ בְּנֵי יִשְׂרָאֵל, I have made the Levites a gift for Aaron and his sons from among the Israelites.

Num 18:7 P עֲבֹדַת מַתָּנָה אֶתֵּן אֶת־כְּהֻנַּתְכֶם, I make your priesthood a gift.

Num 18:19 P כָּל תְּרוּמֹת הַקֳּדָשִׁים אֲשֶׁר יָרִימוּ בְנֵי־יִשְׂרָאֵל לַיהוָה נָתַתִּי לְךָ וּלְבָנֶיךָ וְלִבְנֹתֶיךָ, All the holy offerings that the Israelites present to the Lord I have given to you, together with your sons and daughters, as a perpetual due.

Num 18:21 P הִנֵּה נָתַתִּי כָּל־מַעֲשֵׂר בְּיִשְׂרָאֵל לְנַחֲלָה חֵלֶף עֲבֹדָתָם אֲשֶׁר־הֵם עֹבְדִים אֶת־עֲבֹדַת אֹהֶל מוֹעֵד, To the Levites I have given every tithe in Israel for a possession in return for the service that they perform, the service in the tent of meeting.

Num 18:24 P כִּי אֶת־מַעְשַׂר בְּנֵי־יִשְׂרָאֵל אֲשֶׁר יָרִימוּ לַיהוָה תְּרוּמָה נָתַתִּי לַלְוִיִּם לְנַחֲלָה, Because I have given to the Levites as their portion the tithe of the Israelites, which they set apart as an offering to the Lord.

Num 21:29 JE נָתַן בָּנָיו פְּלֵיטִם וּבְנֹתָיו בַּשְּׁבִית לְמֶלֶךְ אֱמֹרִי סִיחֹון, He has made his sons fugitives, and his daughters captives, to an Amorite king, Sihon.

Num 31:29 P וְנָתַתָּה לְאֶלְעָזָר הַכֹּהֵן תְּרוּמַת יְהוָה, And you make it for Eleazar the priest as an offering to the Lord.

Num 32:5 P יֻתַּן אֶת־הָאָרֶץ הַזֹּאת לַעֲבָדֶיךָ לַאֲחֻזָּה, Let this land be made for your servants a possession.

The following clauses contain שים with a double object construction (direct object + second object with or without ל).

Gen 21:13 E וְגַם אֶת־בֶּן־הָאָמָה לְגוֹי אֲשִׂימֶנּוּ, As for the son of the slave woman, I will make a nation of him.

Gen 21:18 E כִּי־לְגוֹי גָּדוֹל אֲשִׂימֶנּוּ, For I will make a great nation of him.

Genesis 27:37 J הֵן גְּבִיר שַׂמְתִּיו לָךְ, I have already made him your lord.
Gen 28:11 E וַיָּשֶׂם מְרַאֲשֹׁתָיו, And he made (them) his head-place.
Gen 28:18 E אֲשֶׁר־שָׂם מְרַאֲשֹׁתָיו, Which he made his head-place.
Gen 28:18 E וַיָּשֶׂם אֹתָהּ מַצֵּבָה, And he made it a high place.
Gen 28:22 E אֲשֶׁר־שַׂמְתִּי מַצֵּבָה, Which he made a high place.
Gen 45:8 J וַיְשִׂימֵנִי לְאָב לְפַרְעֹה וּלְאָדוֹן לְכָל־בֵּיתוֹ וּמֹשֵׁל בְּכָל־אֶרֶץ מִצְרָיִם, He has made me a father to Pharaoh, and lord of all his house and ruler over all the land of Egypt.
Gen 45:9 J שָׂמַנִי אֱלֹהִים לְאָדוֹן לְכָל־מִצְרָיִם, God has made me lord of all Egypt.
Gen 46:3 E כִּי־לְגוֹי גָּדוֹל אֲשִׂימְךָ שָּׁם, For I will make of you a great nation there.
Gen 47:6 J וְשַׂמְתָּם שָׂרֵי מִקְנֶה עַל־אֲשֶׁר־לִי, And you make them leaders of my livestock.
Gen 47:26 E וַיָּשֶׂם אֹתָהּ יוֹסֵף לְחֹק עַד־הַיּוֹם הַזֶּה עַל־אַדְמַת מִצְרַיִם, So Joseph made it a statute concerning the land of Egypt, and it stands to this day.
Exod 2:14 J מִי שָׂמְךָ לְאִישׁ שַׂר, Who made you a ruler and judge over us?
Exod 14:21 J וַיָּשֶׂם אֶת־הַיָּם לֶחָרָבָה, And he turned the sea into dry land.
Exod 28:12 P וְשַׂמְתָּ אֶת־שְׁתֵּי הָאֲבָנִים עַל כִּתְפֹת הָאֵפֹד אַבְנֵי זִכָּרֹן לִבְנֵי יִשְׂרָאֵל, You make the two stones on the shoulder-pieces of the ephod stones of remembrance for the sons of Israel.
Lev 24:6 P וְשַׂמְתָּ אוֹתָם שְׁתַּיִם מַעֲרָכוֹת, You will make them two rows.

Software

For this research I have used Python and R. Python 3 was used mainly for preprocessing, but in chapter 6, it was also used for the LSTM model. R was used for the visualizations, Random Forest, XGBoost, and k-means clustering.

I used the following packages:

Python

Text-Fabric
Roorda, D., Text-Fabric (version 5.0.3). Python. The Hague: Data Archiving and Networking Services (DANS), 2018a, https://www.dans-labs.github.io/text-fabric. DOI: 10.5281/zenodo.2635046.

R

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org.

You can find the other Python and R packages that were used in the ReadMe file on my GitHub page: https://github.com/MartijnNaaijer/phdthesis.

Literature

Albrecht, C.

1888 "Die Wortstellung im Hebraeischen Nominalsatz II", *Zeitschrift für die Alttestamentliche Wissenschaft*, volume 8, number 1, 249–263.

Andersen, F.I.

1970 *The Hebrew Verbless Clause in the Pentateuch*. Nashville: Abingdon Press.

Baasten, M.F.J.

1997 "Nominal clauses containing a personal pronoun in Qumran Hebrew" in: Muraoka and Elwolde, eds. *1997*, 1–16.

1999 "Nominal clauses with locative and possessive predicates in Qumran Hebrew", in: Muraoka and Elwolde, eds. *1999*, 25–52.

2000 "Existential Clauses in Qumran Hebrew", in: Muraoka and Elwolde, eds. 2000, 1–11.

2006 *The Non-Verbal Clause in Qumran Hebrew*. PhD Dissertation, Leiden University.

Baayen, R.H.

2008 *Analyzing Linguistic Data, A Practical Introduction to Statistics Using* R, Cambridge: Cambridge University Press.

Baayen, R.H., van Rij, J., de Cat, C., and Wood, S.N.

2016 "Autocorrelated errors in experimental data in the language sciences: Some solutions offered by Generalized Additive Mixed Models" in: Speelman et al. *2016*, 49–70.

Baden, J.S.

2012 *The Composition of the Pentateuch, Renewing the Documentary Hypothesis*, New Haven: Yale University Press.

Bakker, D.

In preparation

Syntactic Variation in Classical Hebrew: A Text-Grammatical Approach.

Barmash, ed.
2017 "Symposium: Does Archaic Biblical Hebrew Exist?", *Hebrew Studies* 58, 47–118.

Bartelmus, R.
1982 HYH: *Bedeutung und Funktion eines hebräischen "Allerweltwortes"*, St. Ottilien: Eos Verlag.

Bergey, R.L.
1983 "The Book of Esther—Its Place in the Linguistic Milieu of Post-exilic Biblical Hebrew Prose: A Study in Late Biblical Hebrew", PhD dissertation, Dropsie College for Hebrew and Cognate Learning.

Bijl, E. van de, Kingham, C., Peursen, W.Th. van, and Bhulai, S.
2018 "A Probabilistic Approach to Syntactic Variation in Biblical Hebrew", NIAA Proceedings.

Blau, J.
1993 *A Grammar of Biblical Hebrew*, Wiesbaden: Harrassowitz Verlag, Second, amended edition.
2000 "A Conservative View of the Language of the Dead Sea Scrolls" in: Muraoka and Elwolde, eds. 2000.

Block, D.I. and Schultz, R.L., eds.
2015 *Bind up the Testimony: Explorations in the Genesis of the Book of Isaiah*, Peabody, MA: Hendrickson.

Blum, E.
2016 "The Linguistic Dating of Biblical Texts" in: Gertz et al. 2016.

Brenner-Idan, A., ed.
2014 *Discourse, Dialogue and Debate in the Bible, Essays in Honour of Frank H. Polak*, Sheffield: Sheffield Phoenix Press.

Brockelmann, C.
1956 *Hebräische Syntax*, Neukirchen: Kreis Moers, Verlag der Buchhandlung des Erziehungsvereins.

Chambers, J.K., Trudgill, P. and Schilling-Estes, N., eds.
2004 *The Handbook of Language Variation and Change*, Malden, Oxford: Blackwell.

Cook, J.A., ed.
2002 *Bible and Computer. The Stellenbosch AIBI-6 Conference. 2000-07-17/21*, Leiden: Brill.

Crawley, M.J.
2007 *The R Book*, Chichester: Wiley.

Cross, F.M.
1994 *Canaanite myth and Hebrew epic: Essays in the history of the religion of Israel*. Cambridge, MA: Harvard University Press.

Cross, F.M. and Freedman, D.N.
1997 *Studies in ancient Yahwistic poetry*, 2nd edition, Grand Rapids, Michigan: Eerdmans.

Cryer, F.H.
1994 "The Problem of Dating Biblical Hebrew and the Hebrew of Daniel" in: Jeppesen, K., Nielsen, K. and Rosendal, B., eds., *In the Last Days, On Jewish and Christian Apocalyptic and its Period*, Aarhus: Aarhus University Press.

Davies, P.R.
1995 *In Search of Ancient Israel*, Sheffield: Sheffield Academic Press.

Dixon, R.M.W.
2010 *Basic Linguistic Theory, Grammatical Topics*, volume 2, Oxford: Oxford University Press.

Dixon, R.M.W. and Aikhenvald, A.Y.
2000 *Changing Valency, Case Studies in Transitivity*, Cambridge: Cambridge University Press.

Dozeman, T.B. and Schmid, K.
2006 *A Farewell to the Yahwist: the Composition of the Pentateuch in Recent European interpretation*, Atlanta: Society of Biblical Literature.

Driver, S.R.

1882 "On Some Alleged Linguistic Affinities of the Elohist", *Journal of Philology* 11, 201–236.

1892a *A treatise on the use of the tenses in Hebrew and some other syntactical questions*. 3rd edition. Oxford: Oxford University Press.

1892b *An Introduction to the Literature of the Old Testament*, New York: Charles Scribner's Sons, second edition.

Dyk, J.W.

1984 "'To Be' in Hebrew: Expressions for 'to be' and the Shift in their Usage between Classical and Rabbinical Hebrew", MA Thesis, Vrije Universiteit Amsterdam.

2014 "Traces of Valence Shift in Classical Hebrew" in: Brenner-Idan, ed. 2014, 48–65.

2016 "How do Hebrew Verbs differ? A Flow Chart of Differences" in: Lewis, Salveson, and Turner, eds., 33–51.

Dyk, J.W., Glanz, O., and Oosting, R.

2014 "Analysing Valence Patterns in Biblical Hebrew: Theoretical Questions and Analytic Frameworks", *Journal of Northwest Semitic Languages*, volume 40, no. 1, 43–62.

Dyk, J.W. and Talstra, E.

1999 "Paradigmatic and Syntagmatic Features in Identifying Subject and Predicate in Nominal Clauses" in: Miller, ed. 1999, 133–186.

Eissfeldt, O.

1964 *Einleitung in das Alte Testament unter Einschluß der Apokryphen und Pseudepigraphen sowie der apokryphen- und pseudepigraphenartigen Qumrān-Schriften: Entstehungsgeschichte des Alten Testaments*, Tübingen: Mohr, 1964, 3. neubearbeitete Auflage.

Ewald, H.

1855 Ausführliches *Lehrbuch der Hebräischen Sprache des Alten Bundes*, 6th edition, Göttingen: Verlag der Dieterichschen Buchhandlung.

Fassberg, S.E. and Hurvitz, A., eds.

2006 *Biblical Hebrew in Its Northwest Semitic Setting: Typological and Historical Perspectives*, Jerusalem: Magness Press.

Forbes, A.D.

2016 "The Diachrony Debate, A Tutorial on Methods", *Journal for Semitics*, 25/2, 881–926.

Franka, S.L. and Christiansen, M.H.

2018 'Hierarchical and sequential processing of language A response to: Ding, Melloni, Tian, and Poeppel (2017). Rule-based and word-level statistics-based processing of language: insights from neuroscience. Language, Cognition and Neuroscience.' *Language, Cognition and Neuroscience*, volume 33, No. 9, 1213–1218. https://doi.org/10.1080/23273798.2018.1424347. Consulted on 12 September 2019.

Freedman, D.N., Forbes, A.D., and Andersen, F.I., eds.

1992 *Studies in Hebrew and Aramaic Orthography*, Biblical and Judaic Studies, volume 2, Winona Lake: Eisenbrauns.

Geiger, G.

2012 *Das hebräische Partizip in den Texten aus der judäischen Wüste*, Leiden: Brill.

Gertz, J.C. et al., eds.

2016 *The Formation of the Pentateuch, Bridging the Academic Cultures of Europe, Israel and North America*, Forschungen zum Alten Testament 111, Tübingen: Mohr Verlag.

Gesenius, W.

1815 *Geschichte der hebräischen Sprache und Schrift*, Leipzig: F.C.W. Vogel.

Glanz, O., Oosting, R., and Dyk, J.W.

2015 "Valence Patterns in Biblical Hebrew: Classical Philology and Linguistic Patterns", *Journal of Northwest Semitic Languages*, 41/2.

Goldenberg, G.

2006 "Comments on 'Three Approaches to the Tripartite Nominal Clause' in Syriac by Wido van Peursen" in: Van Keulen and Van Peursen, eds. 2006, 175–184.

Goodwin, D.W.

1969 *Text-restoration Methods in Contemporary U.S.A. Biblical Scholarship*, Naples: Istituto Orientale di Napoli.

Gries, S.T.
2006 "Exploring variability within and between corpora: some methodological considerations", *Corpora* 1 (2), 109–151.

Hardmeier, C., Syring, W.D., Range, J.D., and Talstra, E., eds.
2000 *Ad Fontes! Quellen Erfassen - Lesen - Deuten, Was ist Computerphilologie? Ansatzpunkte und Methodologie—Instrumente und Praxis*, APPLICATIO 15, Amsterdam: VU University Press.

Hasselbach, R. and Pat-El, N., eds.
2012 *Language and Nature, Papers Presented to John Huehnergard on the Occasion of his 65th Birthday*, Chicago: The Oriental Institute of the University of Chicago.

Hastie, T.J. and Tibshirani, R.J.
1990 *Generalized Additive Models*, Chapman & Hall.

Hecke, P. van
2007 "Constituent Order in Existential Clauses" in: Joosten and Rey, eds. 2007, 61–78.

Hendel, R. and Joosten, J.
2018 *How Old is the Hebrew Bible? A Linguistic, Textual, and Historical Study*, New Haven: Yale University Press.

Henry, A.
2004 "Variation and Syntactic Theory" in: Chambers, Trudgill, and Schilling-Estes, eds., 267–282.

Hoftijzer, J.
1973 "The Nominal Clause Reconsidered", *Vetus Testamentum* 23, 446–510.

Holmstedt, R. and Jones, A.R.
2014 "The Pronoun in Tripartite Verbless Clauses in Biblical Hebrew: Resumption for Left-dislocation or Pronominal Copula?", *Journal of Semitic Studies*, LIX/1, 53–89.

Hornkohl, A.D.
2013 "Biblical Hebrew: Periodization" in: Khan, ed. 2013.

2014 *Ancient Hebrew Periodization and the Language of the Book of Jeremiah: the Case for a Sixth-Century Date of Composition*, Leiden: Brill.

Hudson, R.A.

1996 *Sociolinguistics*. Cambridge: Cambridge University Press.

Hurvitz, A.

1972 "Summary of the article: 'Linguistic Observations on the Priestly Term *'edah* and the Language of P'", Immanuel 1, 21–23.

1973 "Linguistic Criteria for Dating Problematic Biblical Texts", *Hebrew Abstracts* 14, 74–79.

1974 "The Date of the Prose-Tale of Job Linguistically Reconsidered", *Harvard Theological. Review* 67, 17–34.

1982 *A Linguistic Study of the Relationship between the Priestly Source and the Book of Ezekiel: A New Approach to an Old Problem*, CahRB 20, Paris: J. Gabalda.

2000a "Once Again: The Linguistic Profile of the Priestly Material in the Pentateuch and its Historical Age, A Response to J. Blenkinsopp", *Zeitschrift für die Alttestamentliche Wissenschaft* 112, 180–191.

2000b "Was QH a Spoken Language? On Some Recent Views and Positions: Comments" in: Muraoka and Elwolde, eds. 2000, 110–114.

2012 "The 'Linguistic Dating of Biblical Texts': Comments on Methodological Guidelines and Philological Procedures" in: Miller-Naudé and Zevit, eds. 2012, 265–280.

Hurvitz, A., in collaboration with Gottlieb, L., Hornkohl, A., and Mastéy, E.

2014 *A Concise Lexicon of Late Biblical Hebrew, Linguistic Innovations in the Language of the Second Temple Period*, Leiden: Brill.

Jacobs, J.T.

2018 *Statistics, Linguistics and the "Biblical" Dead Sea Scrolls*, Oxford: Oxford University Press.

James, G., Witten, D., Hastie, T., and Tibshirani, R.

2013 *An Introduction to Statistical Learning with Applications in* R, Heidelberg: Springer.

Jongeling, K., Murre-van den Berg, H.L., and Van Rompay, L., eds.

1991 *Studies in Hebrew and Aramaic Syntax, Presented to Professor J. Hoftijzer on the Occasion of his Sixty-Fifth Birthday*, Leiden: Brill.

Joosten, J.

2005 "The Distinction Between Classical and Late Biblical Hebrew as Reflected in Syntax", *Hebrew Studies 26*, 327–339.

Joosten, J. and Rey, J.S., eds.

2007 *Conservatism and Innovation in the Hebrew Language of the Hellenistic Period*, Leiden: Brill.

Joüon, P.

2006 *A Grammar of Biblical Hebrew*. Translated and revised by T. Muraoka. Subsidia Biblica 27. Second ed. Rome: Pontifical Biblical Institute.

Kaajan, M.E.

In preparation

Syntactic Variation in Non-Verbal Phrase Structure in Biblical Hebrew. PhD thesis, Vrije Universiteit Amsterdam.

"Main and Subordinate Clauses in Biblical Hebrew".

Kalkman, G.J.

2015 "Verbal Forms in Biblical Hebrew Poetry: Poetic Freedom or Linguistic System?", Unpublished PhD thesis, Vrije Universiteit Amsterdam.

Kautzsch, E., ed. and rev.

1910 *Gesenius' Hebrew Grammar*. Translated and revised by A.E. Cowley. 2nd ed., Oxford: Clarendon.

Kearns, M. and Valiant, L.G.

1989 "Cryptographic limitations on learning Boolean formulae and finite automata", *Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing*, New York: ACM Press, 433–444.

Kelleher, J.D., Mac Namee, B., and D'Arcy, A.

2015 *Fundamentals of Machine Learning and Predictive Data Analytics, Algorithms, Worked Examples and Case Studies*, Cambridge, MA: The MIT Press, 2015.

Keulen, P.S.F. van, and Peursen, W.Th. van, eds.

2006 *Corpus Linguistics and Textual History. A Computer-Assisted Interdisciplinary Approach to the Peshitta*, SSN 48, Assen: Van Gorcum.

Khan, G.

2006 "Some aspects of the copula in North West Semitic" in: Fassberg and Hurvitz, eds., 155–176.

Khan, G., ed.

2013 *Encyclopedia of Hebrew Language and Linguistics Online*, Leiden: Brill.

Kim, D.H.

2013 *Early Biblical Hebrew, Late Biblical Hebrew, and Linguistic Variability, A Sociolinguistic Evaluation of the Linguistic Dating of Biblical Texts*, Leiden: Brill.

Knauf, E.A.

1990 "War 'Biblisch-Hebräisch' eine Sprache?", *Zeitschrift für Althebraistik*, 3, 11–23.

Kropat, A.

1909 *Die Syntax des Autors der Chronik verglichen mit der seiner Quellen*, BZAW, 16, Giessen: A. Töpelmann.

Kuhn, M., and Johnson, K.,

2013 *Applied Predictive Modeling*, Heidelberg: Springer Verlag.

Kutscher, E.Y.

1974 *The Language and Linguistic Background of the Isaiah Scroll (1QIsaa)*, Studies on the Texts of the Desert of Judah, 6. Leiden: Brill.

1982 *A History of the Hebrew Language*, Jerusalem: Magnes Press.

Landes, G.M.

1992 "Review of J.M. Sasson 1990", Journal of Biblical Literature 111, 130–134.

Levshina, N.

2015 *How to Do Linguistics with R, Data Exploration and Statistical Analysis*, Amsterdam: John Benjamins.

Lewis, T.M., Salveson, A.G., and Turner, B., eds.

2016 *Contemporary Examinations of Classical Languages: Valency, Lexicography, Grammar* (PLAL; No. 8). Piscataway: Gorgias Press, 2016.

Liaw, A. and Wiener, M.
2002 'Classification and Regression by randomForest', R *News* 2(3), 18–22.

Lipschits, O. and Oeming, M., eds.
2006 *Judah and the Judeans in the Persian Period*, Winona Lake: Eisenbrauns.

Malessa, M.
2006 *Untersuchungen zur verbalen Valenz im biblischen Hebräisch*, SSN 49, Assen: van Gorcum, 2006.

Mandell, A.
2013 "Archaic Biblical Hebrew" in: Khan, ed. 2013.

McEnery, T. and Hardie, A.
2012 *Corpus Linguistics, Method, Theory and Practice*, Cambridge: Cambridge University Press.

Merwe, C.H.J. van der
1997 "An Overview of Hebrew Narrative Syntax" in: van Wolde 1997.

Merwe, C.H.J. van der, Naudé, J.A., and Kroeze, J.H.
2004 *A Biblical Hebrew Reference Grammar*, London: Continuum, (reprint).

Mikolov, T., Chen, K., Corrado, G., and Dean, J.
2013 'Efficient Estimation of Word Representations in Vector Space', arXiv:1301.3781.

Miller, C.L., ed.
1999 *The Verbless Clause in Biblical Hebrew: Linguistic Approaches*, Winona Lake, Eisenbrauns, 133–185.

Miller-Naudé, C. and Zevit, Z., eds.
2012 *Diachrony in Biblical Hebrew*, Linguistic Studies in Ancient West Semitic 8, Winona Lake: Eisenbrauns.

Morag, Sh.
1988 'Qumran Hebrew, Some Typological Observations', Vetus Testamentum XXXVIII, 2, 148–164.

Moshavi, A., and Notarius, T., eds.

2017 *Advances in Biblical Hebrew Linguistics, Data, Methods, and Analyses*, Winona Lake: Eisenbrauns.

Muraoka, T.

1985 *Emphatic Words and Structures in Biblical Hebrew*, Jerusalem: The Magness Press, Leiden, Brill.

1999 "The Participle in Qumran Hebrew with Special Reference to its Periphrastic Use" in: Muraoka and Elwolde 1999, 188–204.

2013 "Existential: Biblical Hebrew" in: Khan, ed. 2013.

Muraoka, T. and Elwolde, J.F., eds.

1997 *The Hebrew of the Dead Sea Scrolls & Ben Sira, Proceedings of a Symposium Held at Leiden University, 11–14 December 1995*, Leiden: Brill.

1999 *Sirach, Scrolls and Sages, Proceedings of a Second International Symposium on the Hebrew of the Dead Sea Scrolls, Ben Sira, & the Mishnah, Held at Leiden University, 15–17 December 1997*, Leiden: Brill.

2000 *Diggers at the Well, Proceedings of a Third International Symposium of the Hebrew of the Dead Sea Scrolls and Ben Sira*, Studies on the Texts of the Desert of Judah, 36, Leiden: Brill.

Niccacci, A.

1990 *The Syntax of the Verb in Classical Hebrew Prose*, Sheffield: JSOT Press.

Nixon, J.S., Van Rij, J., Mok, P., Baayen, R.H., and Chen, Y.

2016 "The Temporal Dynamics of Perceptual Uncertainty: Eye Movement Evidence from Cantonese Segment and Tone Perception", *Journal of Memory and Language*, volume 90, 103–125.

Noegel, S.B. and Rendsburg, G.

2009 *Solomon's Vineyard: Literary and Linguistic Studies in the Song of Songs*, SBL Ancient Israel and Its Literature, 1, Atlanta: Society of Biblical Literature.

Notarius, T.

2013 *The Verb in Archaic Biblical Poetry, A Discursive, Typological, and Historical Investigation of the Tense System*, Leiden: Brill.

2015 Review of Vern (2011), *Journal of Semitic Studies* 60, 244–248.

Noth, M.

1959 *Das Zweite Buch Mose: Exodus*, Göttingen: Vandenhoeck & Ruprecht.

1960 *Überlieferungsgeschichte des Pentateuch*, Darmstadt: Wissenschaftliche Buchgesellschaft, 2. Auflage.

Oosting, R.

2011 *The Role of Zion/Jerusalem in Isaiah 40–55: A Corpus-Linguistic Approach*, Ph.D. dissertation, VU University Amsterdam, 2011.

2016 "Computer-Assisted Analysis of Old Testament Texts: The Contribution of the wivu to Old Testament Scholarship" in: Spronk, ed. 2016.

Oosting, R., and Dyk, J.W.

2017 "Valence Patterns in Motion Verbs: Syntax, Semantics, and Linguistic Variation", *Journal of Northwest Semitic Languages*, 43/1.

Paul, S.

2012 "Signs of Late Biblical Hebrew in Isaiah 40–66" in: Miller-Naudé and Zevit, eds. 2012, 293–300.

Pérez Fernandez, M.P.

1999 *An Introductory Grammar of Rabbinic Hebrew*, Leiden: Brill.

Peursen, W.Th. van

2004 *The Verbal System in the Hebrew Text of Ben Sira*, Leiden: Brill.

2006 "Three Approaches to the Tripartite Nominal Clause in Classical Syriac" in: van Keulen and van Peursen, eds. 2006, 157–173.

2007 *Language and Interpretation in the Syriac Text of Ben Sira: a Comparative Linguistic and Literary Study*, Leiden: Brill.

Polak, F.H.

2003 "Style Is More than a Person, Literary Culture and the Distinction Between Written and Oral Narrative" in: Young, ed. 2003, 38–103.

2006a "Sociolinguistcs: A Key to the Typology and the Social Background of Biblical Tradition", *Hebrew Studies* 47, 115–162.

2006b "Sociolinguistics and the Judean Speech Community in the Achaemenid Empire". In: Lipschits and Oeming, ed. 2006, 589–628.

Polzin, R.

1976 *Late Biblical Hebrew: Toward an Historical Typology of Biblical Hebrew Prose*, Missoula: Scholars Press, 1976.

Qimron, E.

2000 "The Nature of DSS Hebrew and its Relation to BH and RH" in: Muraoka and Elwolde, eds. 2000.

2008 *The Hebrew of the Dead Sea Scrolls*, Harvard Semitic Studies 29, Winona Lake, IN: Eisenbrauns (paperback edition).

Rendsburg, G.

1990a *Diglossia in Ancient Hebrew*. American Oriental Series 72, New Haven: American Oriental Society.

1990b *Linguistic Evidence for the Northern Origin of Selected Psalms*, Society of Biblical Literature Monograph Series 43, Atlanta: Scholars Press.

1991 "The Strata of Biblical Hebrew", JNSL XVII, 81–99.

2002a *Israelian Hebrew in the Book of Kings*, Occasional Publications of the Department of Near Eastern Studies and the Program of Jewish Studies, Cornell University 5; Bethesda, MD: CDL Press.

2002b "Some False Leads in the Identification of Late Hebrew Texts: The Case of Genesis 24 and 1 Samuel 2:27–36", *Journal of Biblical Literature* 121, 23–46.

2003a 'A Comprehensive Guide to Israelian Hebrew: Grammar and Lexicon', *Orient*, 38, 5–35.

2003b 'Hurvitz Redux: On the Continued Scholarly Inattention to a Simple Principle of Hebrew Philology' in: Young, ed. 2003, 104–128.

2012a "Northern Hebrew Through Time, From the Song of Deborah to the Mishnah" in: Miller-Naudé and Zevit, eds. 2012.

2012b "Late Biblical Hebrew in the Book of Haggai" in: Hasselbach and Pat-El, eds. 2012.

Rezetko, R.

2003 "Dating Biblical Hebrew, Evidence from Samuel-Kings" in: Young, ed. 2003, 215–250.

2013 "The Qumran Scrolls of the Book of Judges, Literary Formation, Textual Criticism and Historical Linguistics", *Journal of Hebrew Scriptures Online*, volume 13, article 2.

Rezetko, R. and Naaijer. M.

2016a Review of Huvitz et al. (2014) in: *Journal of Hebrew Scriptures*, volume 16.

2016b "An Alternative Approach to the Lexicon of Late Biblical Hebrew", volume 16, article 1, *Journal of Hebrew Scriptures*.

Rezetko, R. and Young, I.

2014 *Historical Linguistics and Biblical Hebrew, Steps Toward an Integrated Approach*, Atlanta: SBL Press, 2014.

2019 "Currents in the Historical Linguistics and Linguistic Dating of the Hebrew Bible: Report on the State of Research as Reflected in Recent Major Publications", *Hiphil Novum*, volume 5, no 1.

Robertson, D.A.

1972 *Linguistic Evidence in Dating Early Hebrew Poetry*, SBLDS 3, Montana: University of Montana Press.

Rooker, M.F.

1990 *Biblical Hebrew in Transition: The Language of the Book of Ezekiel*, JSOTSup. 90, Sheffield: Sheffield Academic Press.

2015 "Characteristics of the Hebrew of the Recognized Literary Divisions of Isaiah" in: Block and Schultz, eds. 2015.

Roorda, D.

2018 Coding the Hebrew Bible, Research Data Journal for the Humanities and Social Sciences, https://brill.com/view/journals/rdj/aop/article-10.1163-24523666-01000011.xml, consulted on 28-3-2019. DOI: https://doi.org/10.1163/24523666-01000011.

Roorda, D., Kalkman, G., Naaijer, M., and Van Cranenburgh, A.

2014 "LAF-Fabric: a Data Analysis Tool for Linguistic Annotation Framework with an Application to the Hebrew Bible", *Computational Linguistics in the Netherlands Journal*, volume 4, 2014, 105–109.

Sáenz-Badillos, A.

1993 *A History of the Hebrew Language*, Cambridge: Cambridge University Press.

Sappan, R.
1981 *The Typical features of the Syntax of Biblical Poetry in its Classical Period* (Hebrew), Jerusalem: Kiryath Sepher.

Schneider, W. and Grether, O.
1974 *Grammatik des biblischen Hebräisch: ein Lehrbuch*, München: Claudius.

Schoors, A.
2004 *The Preacher Sought to Find Pleasing Words, a Study of the Language of Qoheleth, Part II, Vocabulary*, OLA 143, Leuven: Peeters.

Schüle, A.
2000 *Syntax der Althebräischen Inschriften, ein Beitrag zur Historischen Grammatik des Hebräischen*, Münster: Ugarit-Verlag.

Shin, S.-Y.
2007 "A Lexical Study on the Language of Haggai-Zechariah-Malachi and its place in the History of Biblical Hebrew", PhD thesis, Hebrew University.

Sinclair, C.
1999 "Are Nominal Clauses a Distinct Clausal Type?" in: Miller, ed. 1999.

Speelman, D., Heylen, K., Geeraerts, D., eds.
2016 *Mixed Effects Regression Models in Linguistics*, Berlin: Springer.

Spronk. K., ed.
2016 *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.

Tagliamonte, S., and Baayen, R.H.
2012 "Models, forests and trees of York English: Was/were Variation as a Case Study for Statistical Practice". Online: http://read.psych.uni-potsdam.de/index.php?option=com_content&view=article&id=78:tagliamonte-and-.baayen-2012-models-forests-and-trees-of-york-english-was-were-variation-as-a-case-study-for-statistical-practice&catid=24:publications&Itemid=32. Consulted on 12 August 2017.

Talstra, E.,

1991 "Hebrew Syntax: Clause Types and Clause Hierarchy" in: Jongeling, Murre-van den Berg, and Van Rompay, eds. 1991.

2002 "Computer-assisted linguistic analysis. The Hebrew Database used in Quest.2" in: Cook, ed. 2002.

Talstra, E., and Sikkel, C.

2000 "Genese Und Kategorienentwicklung Der WIVU-Datenbank" in: Hardmeier, Syring, Range, and Talstra 2000.

Tov, E.

2012 *Textual Criticism of the Hebrew Bible*, Minneapolis: Fortress Press, third edition.

Verheij, A.

1994 *Grammatica Digitalis 1—The Morphological Code in the "Werkgroep Informatica" Computer Text of the Hebrew Bible*, Amsterdam: VU University Press.

Vern, R.C.

2011 *Dating Archaic Biblical Hebrew Poetry, A Critique of the Linguistic Arguments, Piscataway*, NJ: Gorgias Press.

Waltke, B.K., and O'Connor, M.

1990 *An Introduction to Biblical Hebrew Syntax*, Winona Lake: Eisenbrauns.

Wieling, W., Tomaschek, F., Arnold, D., Tiede, M., Bröker, F., Thiele, S., Wood, S.N., and Baayen, R.H.

2016 "Investigating dialectal differences using articulography", Journal of Phonetics, 59, 122–143.

Wellhausen, J.

1883 *Prolegomena zur Geschichte Israels*, Berlin: Druck und verlag von G. Reimer.

Winther-Nielsen, N.

2017 "Give and the meaning of נתן nātan in Genesis" in: Moshavi and Notarius 2017, 363–385.

Wolde, E.J. van, ed.

1997 *Narrative Syntax and the Hebrew Bible: Papers of the Tilburg Conference 1996*, Leiden: Brill.

Wood, S.

2017 *Generalized Additive Models: An Introduction with* R, CRC Press, Second Edition.

Wright, R.M.

2005 *Linguistic Evidence for the Pre-Exilic Date of the Yahwistic Source*, LHBOTS, 419, New York, London: Continuum.

Young, I.

1993 *Diversity in Pre-Exilic Hebrew*, Tübingen: Mohr Verlag.

2003 "Late Biblical Hebrew and Hebrew Inscriptions" in: Young, ed. 2003, 276–311.

2008 "Late Biblical Hebrew and the Qumran Pesher Habakkuk", *Journal of Hebrew Scriptures*, volume 8, article 25, 38 pages.

2009 "Is the Prose-Tale of Job in Late Biblical Hebrew", *Vetus Testamentum* 59, 606–629.

Young, I., ed.

2003 *Biblical Hebrew: Studies in Chronology and Typology*, London: T & T Clark International, 2003.

Young, I., Rezetko, R., and Ehrensvärd, M.

2008 *Linguistic Dating of Biblical Texts*, 2 volumes, London: Equinox Publishing.

Summary

The goal of this thesis is to investigate the linguistic, literary, and historical background of syntactic variation in Biblical Hebrew (BH) from a quantitative perspective. First, the thesis deals with variation in expressions for "to be". The four ways for expressing "to be" that are discussed are clauses with היה, bipartite verbless clauses, clauses with the particle יש, and tripartite verbless clauses. Then, verbal valence is studied, particularly valence patterns of the polyvalent verbs נתן and שים. The focus is on double object constructions. Finally, the variation between Early Biblical Hebrew (EBH) and Late Biblical Hebrew (LBH) is studied in quoted speech and narrative by classifying clauses as EBH or LBH, based on their clause structure. The clauses under consideration are extracted from the EBH and LBH books and the books of Jonah, Ruth, and the prose tale of Job.

Since the beginning of critical scholarship on the Hebrew Bible in the early 19th century, there has been interest in linguistic variation in BH and its background. It was recognized by scholars such as Gesenius and Driver that there was an early and a late variety of the language. In the 20th century, Hurvitz developed a method for dating texts of unknown date based on linguistic characteristics. For Hurvitz, language is a more objective criterion for dating texts than the historical, literary or theological background of a text. In the late 20th and early 21st century a debate arose as to the extent and background of linguistic variation in Biblical Hebrew, and it was questioned whether Biblical Hebrew texts could be dated linguistically. This debate is the background of the "Syntactic Variation Project" of the Vrije Universiteit Amsterdam, of which this research is a subproject.

The dataset that is used throughout this research is that of the Hebrew Bible as encoded in the Eep Talstra Centre for Bible and Computer (ETCBC) database. Next to the Hebrew Bible, a number of extrabiblical Hebrew texts are used in the research as benchmarks.

An important goal of this research is to show the value of quantifying linguistic variation. With multivariate statistical techniques it is possible to investigate the relationships between more than two variables in a single analysis, and to process large quantities of relevant data. Processing of the data was done with Text-Fabric. The development of Text-Fabric, an open source package for storage and processing of language data, is an important step in the integration of quantitative research, biblical studies, and the open science movement.

In the Syntactic Variation project, syntactic variation is interpreted in light of four main variables. These are genre, discourse type, language phase, and main and subordinate clauses.

In this thesis, both statistical and machine learning approaches are used to find out how linguistic variation is conditioned. A Generalized Additive Mixed Model is used for investigating variation in clauses with and without the verb היה with a subject and predicate complement. A machine learning approach is chosen for investigating variation between bipartite clauses with and without יש, and bipartite and tripartite clauses. The techniques used here are Random Forest and Extreme Gradient Boosting (XGBoost). A Long Short-Term Memory network is used for predicting the language phase of clauses in quoted speech and narrative sections. Along with these approaches, a variety of explorative visualizations are used for clarifying patterns in the data.

The statistical analysis of the use of היה in clauses with a subject and predicate complement has shown various tendencies in the data. Some of these confirm the traditional idea that this verb adds TAM to the clause. These are, for instance, the increased use of היה in main clauses, in clauses with a mother that is not a verbless clause and in clauses with a time phrase. היה is used less if the clause contains a question or interjection phrase, which occur relatively often in clauses in quoted speech. In quoted speech in general, the use of היה does not differ significantly from in narrative texts. There is a significantly lower use of היה in poetry than in prose and prophecy, perhaps because the role of time is less predominate in poetry.

היה seems to be used to give structure to the clause, because longer clauses tend to contain היה more often. Also, there is an increased use of היה in clauses in which the predicate complement is a PP, which seems to confirm the idea that there is semantic variation in the use of היה, indicating that היה has a broader function than only adding TAM.

The verb היה is used significantly more often in EBH than in LBH, in clauses containing a subject and predicate complement. This lower frequency in LBH is not observable in QH or RH, so LBH has a distinct position here.

The present research confirms the interpretation that the particle יש puts emphasis on the clause. The particle is relatively rare, but in contrast to clauses with a similar structure (clauses with an indefinite subject and a PP predicate complement) without the particle, it occurs predominantly in clauses in quoted speech sections. In the analysis with XGBoost, quoted speech is the most important predictor for the presence of יש. Of the different language phases, the particle occurs most often in RH, which is generally associated with spoken language. Like in clauses with היה, longer clauses tend to contain יש more often than do short clauses.

The research on the tripartite verbless clause suggests a non-copular interpretation of the tripartite clause in BH. An important argument is that the structure is relatively rare (fewer than 200 cases in the MT, which is only a small fraction of the total amount of relevant data), one would expect more cases if it were used as a copula. Another clear sign is that in the analysis using XGBoost, quoted speech is the most important predictor for the use of the tripartite clause, also for a situation in which the fronted subject is resumed and emphasized, quoted speech is the natural environment. The second most important predictor for the tripartite clause is whether the clause is an argument clause. Most of these are object clauses, occurring in quoted speech.

Of course, it is theoretically possible that in quoted speech the pronoun functions as a copula, and that the increased use of this copula in quoted speech shows something about the difference between spoken and written Hebrew, but I prefer to use the simplest explanation of the phenomenon. Resumption also occurs in constructions other than tripartite clauses, so with this explanation no new grammatical category needs to be invented or borrowed for BH.

Both clauses with יש and tripartite verbless clauses have a preference for longer subjects. It is possible that יש or the pronoun gives structure to the clause, as was suggested by Driver, in the case of the tripartite clause. On the other hand, conclusions should be drawn with care, because clauses with יש are relatively rare.

It has become clear that there is a variety of factors influencing the use of היה, יש, and the pronominal copula, and these factors have a varied background. Also, studying these together in one analysis has shown the relative importance of these factors.

The variation in the distribution of נתן and שים was studied exploratively. The focus was on double object constructions with and without ל, in which the verbs have (more or less) the same meaning, namely: "to make object_1 to be object_2". Double object constructions with נתן and שים occur relatively often in the Major Prophets: a high frequency with these constructions are found in the books of Isaiah, Jeremiah, and Ezekiel. There does not seem to be much variation in the frequency of double object constructions between the main levels of the discourse environment (N and Q), and between main and subordinate clauses.

Within the prophetic books there is substantial variation of the preference for one of the verbs. In the book of Isaiah, there is a preference for using שים, whereas Ezekiel has a strong preference for using נתן. In the poetic books the Psalms have an equal use of both verbs, but Job and the Song of Songs prefer שים. Between the prose books there is substantial variation in the preference for one of the two alternative verbs. Double object constructions with these verbs are nearly absent in the LBH books of

Esther, Daniel, Ezra, and Nehemia, but it is common in the book of Chronicles, which has a strong preference for the use of נתן. This is not only reflected in the absolute frequencies of נתן and שים in double object constructions (15 times with נתן and once with שים), but also in parallel passages. In various parallels Chronicles uses the verb נתן, where the parallel uses a different verb (e.g., שים) or a different construction with the same verb.

In the Pentateuch, on the other hand, most books have a mixed profile. Deuteronomy has a strong preference for using נתן, and the same is true for the P source. J and E, however, have a preference for using שים. The Former Prophets have a mixed profile. Samuel has a preference for שים, but Kings prefers נתן.

A rarer construction is formed by נתן and שים with double object, of which the second object is introduced by ב. The meaning of the clause is "to make object_1 like object_2". Although it is rare or absent in most books, the distribution of נתן and שים with this construction follows the same patterns as the other double object constructions.

Thus, even though the verbs נתן and שים are synonyms in double object constructions, they are not distributed evenly throughout the Hebrew Bible. Various texts and books have a strong preference for one of the two verbs. The most notable are the assumed sources of the Pentateuch and the book of Chronicles. On the basis of the evidence in the MT, one can conclude that different texts with different backgrounds used a different verb, but it is possible that one verb does not necessarily relate to one specific background. If it is accepted that P is exilic/post-exilic, it is possible that the use of נתן with double object constructions is characteristic of post-exilic Hebrew, because it is used predominantly in Ezekiel, and nearly exclusively in Chronicles.

Linguistic dating is a discipline for which predictive modeling offers a natural solution. A model is trained on features from two subcorpora, EBH and LBH, and predictions are made on unseen data, a text of unknown date, to find out whether the text's language is more similar to EBH or LBH. In chapter 6, this approach was used to find out what the linguistic relationships are between the separate EBH and LBH books and the relationships between those books and the books of Jonah, Ruth, and the prose tale of Job. The results are compared with the results of traditional linguistic dating.

In this research, an LSTM model was used to classify clauses from Jonah, the prose tale of Job, and Ruth. With the LSTM network, it is possible to model sequence data. Modeling sequences without the need to extract specific features from them is an important step forward in language modeling, because LSTM models are able to detect long term dependencies in sequences.

Two analyses were done, one on phrase level, in which clauses are represented as sequences of phrase functions. The other analysis is done on word level, in which clauses are represented as sequences of parts of speech. Generally, the traditional distinction between EBH and LBH is visible in this analysis. If the data are clustered in 2 clusters, one cluster contains mainly the EBH books, whereas the other cluster contains most of the LBH books, but there are some EBH and LBH books falling "between the clusters", indicating that there is no sharp distinction between EBH and LBH. This is an important result, because it confirms the traditional distinction between EBH and LBH. On the other hand, it shows that it is difficult to classify the books of undisputed date in the cluster in which they are supposed to belong. Also, results may vary slightly if other features are chosen as input for the model.

In some studies on linguistic dating, Jonah, the prose tale of Job, and Ruth are considered to be written in LBH. This is based on research in which distinct features are selected that are considered to be typical of LBH. This traditional approach has various difficulties attached to it, such as the weight that needs to be given to each of the features. This and various other problems are solved automatically by the LSTM network. Features contribute to the model as far as they contribute to its predictive power. The result of the analysis is that the clause structure of Jonah, the prose tale of Job, and Ruth are basically that of EBH. Of course, this is based on the way clauses were represented in the analysis, as sequence of phrase functions and as sequence of parts of speech, but this kind of analysis can be extended easily, depending on one's own preferences for features.

How can these results be explained in light of the traditional linguistic dating approach? In traditional research, a number of features is selected that links a text with the core LBH books. However, as will be shown in chapter 2 in the example about Second and Third Isaiah, it is not always clear how features are selected and to what extent late features are representative of LBH in general. With a large-scale quantitative approach, in which many features are taken into account and weighed, this problem is avoided. Linguistic dating is not only based on the idea that texts share linguistic characteristics, but there is a range of assumptions underlying it. There is no clear empirical evidence that linguistic dating of biblical texts works, which makes it a method without any validation. In my opinion, it is more fruitful to analyze linguistic variation from a more descriptive perspective. It is not unlikely that a multidisciplinary approach, in which linguistic, literary and textual history are taken into account may once lead to more insight in the linguistic history of Biblical Hebrew.

This research has shown that linguistic variation in BH has a varied background, being partly literary, partly linguistic, and partly historical. There is variation in the use of היה between different genres. Also, double object constructions with the verbs נתן and שים are relatively frequent in prophecy, especially in the Major Prophets, and various books and redactional layers have a preference for one of the verbs with a double object construction.

A substantial part of the variation studied in this research has a linguistic background. יש and the tripartite verbless clause add emphasis, היה adds TAM, the length of a clause and also the length of the subject of a clause influence the use of היה, יש, and the tripartite verbless clause. Also, there is a clear difference in the use of יש and the tripartite verbless clause between narrative and quoted speech. In chapter 6, it was shown that clauses in quoted speech are more homogeneous in variation between EBH and LBH than clauses in narrative.

Finally, part of the background of the variation studied in this research is historical. In chapter 6 it was shown that there is variation between EBH and LBH, although this variation is not so strong that it is possible to draw a clear border between them. Also, it seems that נתן with a double object construction is preferred in late literature.

All in all, in BH, linguistic variation is conditioned by a variety of factors, even in the case of a single linguistic feature. Quantitative, multivariate techniques are an important addition to the toolset for investigating linguistic variation in BH. It is to be expected, that in the near future, quantitative approaches will rapidly become more influential in Biblical Studies.

Curriculum Vitae

Martijn Naaijer was born on December 2, 1975 in Vlissingen, The Netherlands. After graduation from the Christelijke Scholengemeenshap Walcheren in Middelburg in 1994, he started with Plant Breeding at Wageningen University. After a year, after obtaining the propaedeutics, he changed to Molecular Life Sciences at the same university.

After obtaining his BSc in 1998, he started the MSc programme of Molecular Life Sciences, with a focus on bio-organic chemistry. During the course of this MSc programme his interest in theology grew, and therefore, he started studying Theology at Radboud University Nijmegen.

During this period, he was also appointed as a teacher of physics and chemistry at Penta College CSG (Jacob van Liesveldt in Hellevoetsluis and Blaise Pascal in Spijkenisse), and as a bookseller at the VU Bookstore in Amsterdam.

He obtained his MA degree in 2012. He wrote his MA thesis on the vocabulary of the book of Esther. In this thesis, he started exploring the world of computational linguistics.

In 2013, he started a full time PhD trajectory at the Eep Talstra Centre for Bible and Computer (ETCBC) at Vrije Universiteit Amsterdam, on the NWO-funded project "Does Syntactic Variation reflect Language Change? Tracing Syntactic Diversity in Biblical Hebrew Texts". During his PhD trajectory, he gave several presentations at conferences in the Netherlands, Israel, Poland, and the US. Also, he developed and taught the course "Biblical Studies and Digital Humanities" at Vrije Universiteit Amsterdam, for which he was awarded with the Grassroots prize for educational innovation.

Since 2018, Martijn is working as a data scientist at Bloxs Software in Utrecht, where he a analyzes the dynamics of the real estate market, and develops smart tools for the automatization of the administration of Bloxs' customers.

Set in Trinité, 11 pt
Printed by Ipskamp, Enschede
Typesetting by TAT Zetwerk, Utrecht