Modeling Noun Semantics with Vector Spaces and the ETCBC Hebrew Database

Cody Kingham
Vrije Universiteit Amsterdam, Faculty of Theology, De Boelelaan 1105, 1081 HV
codyakingham@gmail.com

Abstract: Form-to-function Hebrew grammarians have long argued for the priority of syntax over semantics in the linguistic analysis of Biblical Hebrew. For forty years, that philosophy has shaped the model of the Eep Talstra Centre for Bible and Computer (ETCBC) Hebrew database. But recent research has called attention to the interconnectedness of syntax and semantics. Work in verbal valency, in Hebrew, and automatic sense extraction, in natural language processing, are examples of such work. Verbal valency demonstrates how semantics are equally dependent on syntax; methods of automatic sense extraction, such as Hearst patterns, show that world-knowledge is modeled in the language through syntactic relationships. Both of these observations provide a promising foundation for enriching the ETCBC database with form-to-function, semantic data. In this paper, I lay out a new methodology for form-to-function semantics using the ETCBC data. The first section briefly explores the history of syntax and semantic research in form-to-function linguistic studies. In the second section, I provide initial results in using formal structures to extract semantic relations. Those structures are gathered and modeled using semantic vector spaces with syntactic context selection. That data will provide a starting point for a new data package in Text-Fabric format that will enrich the ETCBC (BHSA) with new semantic data.

Keywords: Hebrew linguistics, form-to-function, semantics, natural language processing

Semantics and the ETCBC

The form-to-function approach to Hebrew grammar has now occupied a niche and committed section of Hebraists for over forty years. The method emphasizes the need to describe and analyze the surface structure of the text before interpreting its meaning. For grammar, that means cataloguing grammatical forms and examining their distribution before hypothesizing their function.[1] Talstra, in particular, has used the approach to build the Eep Talstra for Bible and Computer (ETCBC) Hebrew database.[2] The ETCBC restricts its analysis to surface level features by using programs that apply pattern searches and rules-based programs to segment and parse linguistic units. In this way, the database can serve as a tool for discovering language mechanisms rather than simply mirroring the theorist's presuppositions.

[1] As done by Schneider. Wolfgang Schneider, *Grammatik Des Biblischen Hebräisch : Ein Lehrbuch. Völlig Neue Bearb. Der "Hebräischen Grammatik Für Den Akademischen Unterricht* (München: Claudius, 1974), 182–207.

[2] E.g. Eep Talstra, "Text Grammar and Hebrew Bible. I: Elements of a Theory," *BO* 35 (1978): 169–174; Eep Talstra, "Exegesis and the Computer Science: Questions for the Text and Questions for the Computer," *BO* 37, no. 3/4 (1980): 121–128; Eep Talstra, "An Hierarchically Structured Data Base of Biblical Hebrew Texts. The Relationship of Grammar and Encoding," in *Proceedings of the First International Colloquium, Bible and Computer: Interpretation, Heremeneutics, Expertise. Tübingen, 2-3-4 Septembre 1985*, Association Internationale Bible et Informatique (Paris: Champion, 1986), 337–349.

Hand in hand with its focus on formal structure, the ETCBC has stressed the priority of syntax over semantics. That is, before explaining the meaning of a text, one should first seek to describe its formal structure. The "semantic analysis should be performed within the framework set by the syntactic features of a text."[3] Because of this, the database has eschewed many functional, semantic labels. For example, the only semantic information stored for words are minimalistic glosses derived from Koehler and Baumgartner. Likewise, semantic relations between words remain underdeveloped. For instance, in the present implementation of the database, one can easily find all prepositional phrases. But there is no straightforward way of isolating the object of a preposition. To give another example, semantic relationships between phrases are not formally encoded; so if one is to find the predicate of a subject phrase, it has to be found in an indirect way.[4]

These shortcomings both limit the capabilities of future research and open up the ETCBC method to criticism. Word-level semantic information is crucial for research into participant tracking, verbal valency, and the Hebrew verbal system.[5] Without semantic classes for nouns, for instance, participant tracking software has to depend on its human operator for making logical connections between words. One must know what nouns can do what things, and the researcher must theorize, during the analysis, what the text is saying. For verbal valency research, the lack of semantic relations between words leads to difficulties in parsing the various satellites around a verb.[6] There are no formal labels to distinguish classifications such as objects from indirect objects. For research on the verbal system, the lack of verb classes hinders the ability to study the verb form's distribution within semantic classes.[7] In these areas, the gap is being filled by the intuition of the researchers. Yet, this approach is in tension with the ETCBC's own guiding principle of form-to-function, surface level analysis.

The cause for this oversight is a lack of progress in the form-to-function methodology. Most of the ETCBC encoding method was created during the 1980's and 1990's, when computer-

[3] Eep Talstra, "Text Grammar and Hebrew Bible II: Syntax and Semantics," *BO* 34, no. 1/2 (1982): 38.

[4] Though the labels are there for the phrases, there is no link between the phrases themselves except for the indirect connection of being within the same clause.

[5] Recent works illustrating the need can be seen in Reinoud Oosting, *The Role of Zion/Jerusalem in Isaiah 40–55: A Corpus-Linguistic Approach*, vol. 59, Studia Semitica Neerlandica (Leiden: Brill, 2013); Janet Dyk, Oliver Glanz, and Reinoud Oosting, "Analysing Valence Patterns in Biblical Hebrew: Theoretical Questions and Analytic Frameworks," *JNSL* 40, no. 1 (2014): 1–20; Cody Kingham, "Verb in Biblical Hebrew," Data Repository, last modified 2017, accessed March 20, 2018, https://github.com/codykingham/Verb_in_Biblical_Hebrew.

[6] See especially the critiques in Oliver Glanz, Reinoud Oosting, and Janet Dyk, "Valence Patterns in Biblical Hebrew: Classical Philology and Linguistic Patterns," *JNSL* 41, no. 2 (2015): 31–55.

[7] Cook applies a thorough application of semantic classes to examine verbs. Cook does not use statistical or formal criteria, however. This is a need that new ETCBC data might meet. John A. Cook, *Time and the Biblical Hebrew Verb: The Expression of Tense, Aspect, and Modality in Biblical Hebrew*, Linguistic Studies in Ancient West Semitic 7 (Winona Lake: Eisenbrauns, 2012).

assisted linguistic analysis was in its infancy.[8] Since form-to-function grammarians hold that semantic meaning derives from textual structure, and since the database was still in production, semantic categories were passed over in favor of syntactic categories. But now that research has progressed beyond the level of syntax into more interpretive questions, the signs of wear have begun to show. In 1997, towards the end of the center's methodological development, Talstra wrote: "Moreover, I assume the difficulties in identifying all actors in a text and the complications of semantic or pragmatic analysis will make it virtually impossible for a computer programme ever to produce completely correct textual structures."[9] Yet, as any user of Siri or Google well knows, advancements in text analysis have dramatically changed what is possible. Automated tools for named entity recognition, similar to Talstra's participant tracking, are now widely used in the world of Natural Language Processing and Text-Mining.[10] Nowadays, systems are being built that can evaluate logical entailment through ontologies (e.g. Wordnet) and machine learning.[11] Moreover, machine learning has unlocked the ability to uncover all kinds of high level patterns amongst linguistic data that would not be discernible otherwise.

Aside from methodological limitations that have contributed to the lack of developed semantics in the ETCBC, there are technical limitations. Many of the programs used to process the linguistic data are written in Pascal or C++.[12] These low-level programs are not only difficult to grasp by novice programming theologians, they also cannot take advantage of the data science libraries now available in languages like Python or R. The format of data analysis, as well, inherently limits the complexity of semantic information that can be modeled in the database. For instance, the ETCBC represents some word-level semantic information through a unit called the subphrase, which encodes data such as *nomen regens* / *rectum* (construct) relations. But for any given word, there is a limit of three subphrase

[8] Eep Talstra, "On Text and Tools. A Short History of the 'Werkgroep Informatica' (1977-1987)," in *Computer Assisted Analysis of Biblical Texts. Papers Read at Teh Workshop on the Occasion of the Tenth Anniversary of the "Werkgroep Informatica," Faculty of Theology, Vrije Universiteit, Amsterdam*, ed. Eep Talstra (Amsterdam: Free University Press, 1989), 9–28; Reinoud Oosting, "Computer-Assisted Analysis of Old Testament Texts: The Contribution of the WIVU to Old Testament Scholarship," in *The Present State of Old Testament Studies in the Low Countries: A Collection of Old Testament Studies Published on the Occasion of the Seventy-Fifth Anniversary of the Oudtestamentisch Werkgezelschap*, ed. Klaas Spronk, vol. 69, Oudtestamentische Studiën (Leiden: Brill, 2016), 192–209.

[9] Eep Talstra, "A Hierarchy of Clauses in Biblical Hebrew Narrative," in *Narrative Syntax and the Hebrew Bible: Papers of the Tilbug Conference 1996*, ed. Ellen Van Wolde, vol. 29, BibInt (Leiden: Brill, 1997), 92.

[10] Piek Vossen, et al. "NewsReader: Using Knowledge Resources in a Cross-Lingual Reading Machine to Generate More Knowledge from Massive Streams of News," *Knowledge-Based Systems* 110 (2016): 60–85.

[11] Lasha Abzianidze, "A Natural Proof System for Natural Language" (PhD Dissertation, Tilburg Univeristy, 2017), accessed March 20, 2018, https://pure.uvt.nl/portal/files/14858339/Abzianidze_Natural_20_01_2017.pdf.

[12] The ETCBC data creation programs are accessible at https://github.com/ETCBC/data_creation.

relations which can be represented. This is due to the plain-text file, columnar data format in which subphrases are stored.[13]

This paper seeks to initiate methodological development in form-to-function semantics by experimenting with vector spaces as a method of capturing word-level meaning with formal structure. For that task, the experiment utilizes the ETCBC data in Text-Fabric representation. Using the Text-Fabric export function, new features are created on subject and object phrases which allow the head nouns to be easily retrieved. Using the nouns, a semantic vector space is built using syntactic selections, following Padó and Lapata.[14] The results provide the seeds for a future module to the ETCBC data with semantic classes.

Methodology: The Long Way Around

"Texts are not formed by a simple stringing together of words and clauses. They are structured as a sense-unified whole through an interlacing of many and diverse references in addition to the whole clause."[15] In this statement, Schneider demonstrated an awareness of the complexity to semantics that would anticipate advancements in corpus linguistics two decades later.[16] Simply put, the meaning of a text is more than the sum of its parts. Every sense arises from the orientation and direction of a reader through a series of signs that together constitute a unique message. It is not enough, then, to simply describe what a series of words mean and then compile the data. Rather, the interpretation of a text's meaning, and its contained words, is linked to the textual structure that embeds it. In recent years, linguists have begun to challenge the notion that lexicon and syntax are indeed separate at all.[17] Constructions of words and phrases often have irreducible meaning that is greater than the sum of its parts. The words in the phrase "why did the chicken cross the road?" cannot merely be reduced to their component parts, for instance.

Yet Schneider's formulation goes too far in the direction of pure structuralism. As Talstra observes, "word meanings contribute much more to the structure of a text than Schneider seems to allow for."[18] It has been known for some time now that a word's basic meaning

---

[13] The ETCBC data creation process is extensively documented at Cody Kingham, "ETCBC Data Creation," last modified March 3, 2018, accessed March 20, 2018, http://www.etcbc.nl/datacreation/.

[14] Sebastian Padó and Mirella Lapata, "Dependency-Based Construction of Semantic Space Models," *Computational Linguistics* 33, no. 2 (2007): 161–199.

[15] Wolfgang Schneider, *Grammar of Biblical Hebrew*, trans. Randall Mckinion, vol. 1, Studies in Biblical Hebrew (New York: Peter Lang, 2016), 203.

[16] Namely through the developments of construction and pattern grammar. See references below.

[17] Adele E. Goldberg, *Constructions: A Construction Grammar Approach to Argument Structure* (Chicago: University of Chicago Press, 1995); Susan Hunston and Gill Francis, *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*, vol. 4, Studies in Corpus Linguistics (Amsterdam: John Benjamins, 2000).

[18] Talstra, "Syntax and Semantics," 35.

contributes to its combinability with other words.[19] Corpus linguistics, especially, has utilized this basic principle in order to examine and describe the qualities of various words and constructions.[20] The method of discourse analysis in general, especially with regard to its treatment of the verb, has received strong criticism from semantic proponents like Cook.[21] As he rightly argues, the simple diagnosis that the verb form merely indicates communication type does not go far enough in the direction of describing the basic semantic meaning of the verb itself.

The basic meaning of a word, or its generalized collocation preferences, is precisely what is missing from the form-to-function approach to Hebrew grammar, and from the ETCBC database more specifically. By measuring the ability of words to combine with other words, and by grouping together words with similar combinability, the form-to-function method will be able to integrate its structural insights with semantic logic. Do certain verb forms which indicate a narrative situation do so because of their temporal reference? Or do surprising combinations of infrequently collocated words co-occur with certain syntactic structures? These are the kinds of questions one might answer, in addition to participant tracking, valency, and more, with developed semantic data.

How does a form-to-function approach to word-level semantics differ from other approaches? An alternative approach to the semantic gap in the ETCBC data would be to appropriate data which has already been prepared, for instance, by the Semantic Dictionary of Biblical Hebrew.[22] Therein, words are grouped together into objects, events, or relationals, each with nested semantic classes such as "animals" or "substances."[23] This data is likely very useful for some of the research questions already outlined above. However, the approach of manually selecting and curating semantic classes is fundamentally different from the methodology of form-to-function, which depends on explicit indicators and statistical criteria for the identification of function.

---

[19] See for instance J.R. Firth, "A Synopsis of Linguistic Theory, 1930–1995," in *Studies in Linguistic Analysis* (Oxford: Basil Blackwell, 1962), 1–32.

[20] R. Xiao, "Collocation," in *The Cambridge Handbook of English Corpus Linguistics*, ed. D. Biber and R. Reppen (Cambridge: Cambridge University, 2015), 106–124; Anatol Stefanowitsch and Stefan Th. Gries, "Collostructions: Investigating the Interaction of Words and Constructions," *International Journal of Corpus Linguistics* 8, no. 2 (2003): 209–243.

[21] "My central critique of discourse-prominent theory with regard to the BHVS, which has arisen at various points throughout the preceding survey, is that the functional identifications of the verb forms apart from attention to semantics is methodologically problematic. In the most extreme cases, a semantic component of certain verbal forms is denied altogether. More often, however, the theories proceed from discourse-pragmatic functions to forms but fail to persuade because of the numerous exceptions to their function-form correlations. Talstra lauded Schneider's theory as more sound than other discourse theories because he proceeded instead from form to function, but Talstra likewise recognized that Schneider's theory suffered from a neglect of semantics." Cook, *Time*, 171–172.

[22] "A Semantic Dictionary of Biblical Hebrew." Accessed 20 March 2018. http://www.sdbh.org/home-en.html

[23] Reinier de Blois, "Towards a New Dictionary of Biblical Hebrew Based on Semantic Domains" (PhD Dissertation, Vrije Universiteit Amsterdam, 2000).

The difference is significant for three main reasons. First, the trial-and-error process of computationally processing linguistic forms reveals the mechanisms of the language. For instance, instead of manually segmenting clauses, the ETCBC method has developed pattern searching and rule-based programs that make informed suggestions for the divisions. By taking the long way around, the method has effectively built up a clause construction grammar. Similarly, by addressing the problem of word semantics through collocation strengths, the form-to-function method is likely to uncover, along the way, features of noun selection that would otherwise go unnoticed. Second, by computationally discerning word semantics, a lot of useful granular data will have to be generated that can then be reused in higher levels of analysis. For instance, the process of extracting head nouns, described herein, will provide data that can be used later on for other research. Another benefit of preserving each decision made in code is that the decisions and process of each classification are documented and open for challenging. Third, a computational approach is likely to preserve unexpected gaps or idiosyncrasies in the semantic relations portrayed in the Hebrew Bible. An example is, as found in this study, the term שִׁפְחָה "maidservant," a person, which apparently occurs frequently in lists that include livestock. In the cognitive world as presented by the text of the Hebrew Bible, the status of the noun שִׁפְחָה as a person appears unaccented, with other attributes (i.e. servanthood) featured instead. A form-to-function method preserves this unique presentation.

Semantic Vectors Experiment

The proposed way of studying word-level semantics presents a unique challenge: how to capture semantic information through formal features? The Natural Language Processing and Text-Mining fields utilize a few candidate methods. Semantic information is often reflected by certain constructions like the so-called Hearst patterns.[24] An example is the pattern "a noun such as a noun," which indicates a hyper/hyponym relation. But these kinds of patterns seem quite rare in the Hebrew Bible.[25] Another method is a semantic vector space model. This method records word co-occurrences for a given set of target words. The target words are then compared with a similarity metric to identify words with similar senses. Using this approach, one can observe clusters of words with comparable properties that can be further processed with code.

The specific vector space method selected for this experiment is that of Padó and Lapata. In their article, "Dependency-Based Construction of Semantic Space Models," they use syntactic paths to select co-occurrences for a given target word. Through the selection process, co-occurring lexemes can be weighted based on syntactic path lengths. For the purpose of this test, three relations were identified as potentially valuable candidates for relaying semantic data, based on manual inspections of the data: 1) a subject noun to its verb, 2) an object noun to its verb, 3) any noun with another coordinated noun (i.e. *noun and*

[24] Marti Hearst, "Automatic Acquisition of Hyponyms from Large Text Corpora," *Proceedings of the Fourteenth International Conference on Computational Linguistics* (1992); Rion Snow, Daniel Jurafsky, and Andrew Ng, "Learning Syntactic Patterns for Automatic Hypernym Discovery," *Proceeding NIPS'04 Proceedings of the 17th International Conference on Neural Information Processing Systems* (2004): 1297–1304.

[25] See the notebook linked below which these kinds of patterns were sought.

noun).[26] These relations are captured with a tag containing [noun_role] + [verb_stem + verb_lexeme] or [noun_role] + [noun_role]. The basic question sought by these selections is: what kinds of things can certain nouns *do*, can be done *to* them, and be done *alongside* them? The selection process ignores cases where the verb lexeme is simply היה "to be." Nouns are selected within the narrative domain so as to initially reduce noise from poetic uses. Proper names are also excluded. Association measures are then applied to the raw counts to account for irregular distributions of co-occurrences. Two measures were tested alongside one another and yielded similar results. These are the Log-Likelihood approach suggested by Padó and Lapata and Pointwise Mutual Information scores suggested by Levshina.[27] The cosine similarity measure is used to compare the similarities between word co-occurrences.[28]

The experiment uses the Biblical Hebrew data stored in the open-source *Biblia Hebraica Stuttgartensia Amstelodamensis* (BHSA) Text-Fabric representation of the ETCBC data.[29] Text-Fabric stores data about linguistic objects (words, phrases, etc.) in a graph concept.[30] Objects are nodes, data about objects as features on those nodes, and relationships between objects are stored as edges. This representation allows data to be easily manipulated in a Python Jupyter notebook environment. The implementation makes adding features to nodes as simple as feeding a dictionary keyed with node numbers and features as values to a save function.

The BHSA data first had to be processed so that head nouns could be separated from their enclosing phrases. This is because the data contains little explicit embedding information beneath the level of the phrase. Instead, words are grouped into units called subphrases, which are themselves related through edge relationships from one subphrase to another. Relations modeled in subphrases include apposition, *nomen regens/rectum*, and others.[31] In order to extract head nouns, a function had to be created which pulls words not contained within a dependent subphrase.[32] This is complicated by the fact that words can be included in multiple subphrases. Thus, the algorithm has to select all of the subphrases of a given word and check simultaneously for dependent relations. In the case of quantifiers such as כל "all"

[26] See the notebook for the selection process of these features https://github.com/codykingham/semantics/blob/master/2.%20Context%20Selection%20Discovery.ipynb

[27] Padó and Lapata, "Semantic Space Models," 173–174; Natalia Levshina, *How to Do Linguistics with R: Data Exploration and Statistical Analysis* (Amsterdam: John Benjamins, 2015), 326–327.

[28] Levshina, *How to Do Linguistics with R: Data Exploration and Statistical Analysis*, 328–329.

[29] The BHSA data is stored at https://github.com/ETCBC/bhsa

[30] For an introduction to Text-Fabric, see Dirk Roorda, "Text Fabric Wiki," last modified January 27, 2018, accessed March 20, 2018, https://github.com/Dans-labs/text-fabric/wiki.

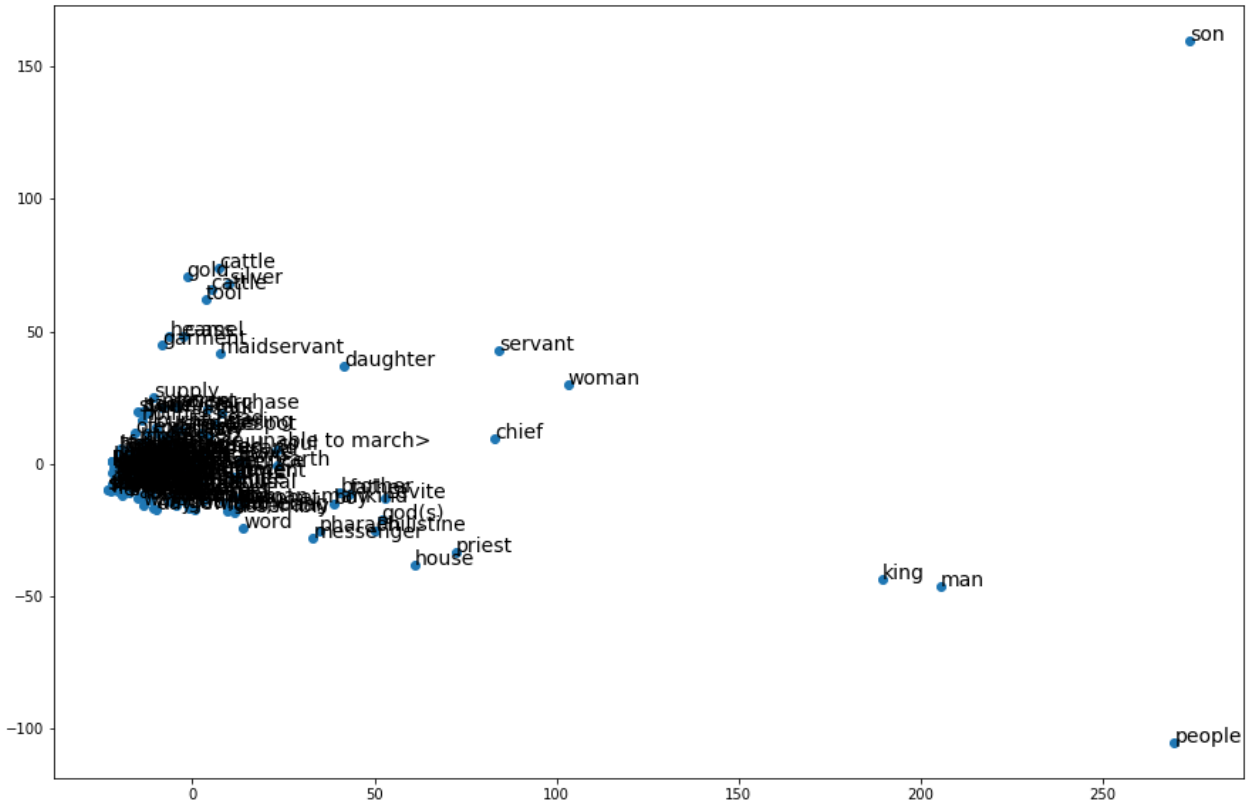
[31] All the subphrase relation features are documented at https://etcbc.github.io/bhsa/features/hebrew/c/rela

[32] The function is developed in the notebook at https://github.com/codykingham/semantics/blob/master/3.%20Context%20Selection%20Development.ipynb

or numbers, the algorithm selects the quantified noun. Using this algorithm, a new set of edge features are exported in TF format to be used as a module alongside the BHSA data. Every phrase node with a valid noun then possesses an edge relation leading to its head noun(s). By calling the edge feature `heads`, a tuple is returned containing the relevant word nodes.

Pictured below are the plotted results using the log-likelihood method.[33] After the non-narrative texts and low frequency results are filtered out, the model contains 189 nouns. These are nouns used only in subject or object phrases. Each dot represents a word. The closer the word in the scatter plot, the more related they are according to the similarity measure. The Jupyter notebook used to generate this plot is linked the footnote.[34]

Figure 1. Semantic Space with Log-Likelihood



The nouns to the far right of the plot are the highest occurring within the dataset, which is why they are so far removed from the other results.[35] One can note already the cluster between 50 and 100 on the y-axis containing זהב "gold", כסף "silver," צאן "cattle," and שִׁפְחָה "maidservant." Closer views of the clusters on the left are depicted below.

[33] The multidimensional vectors are reduced using Principal Component Analysis. http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

[34] https://github.com/codykingham/semantics/blob/master/4.%20Semantic%20Space%20Construction.ipynb

[35] A larger occurrence frequency means that more diverse relationships are contained in the observations.

Figure 2. Semantic Space with Log-Likelihood, Large Cluster

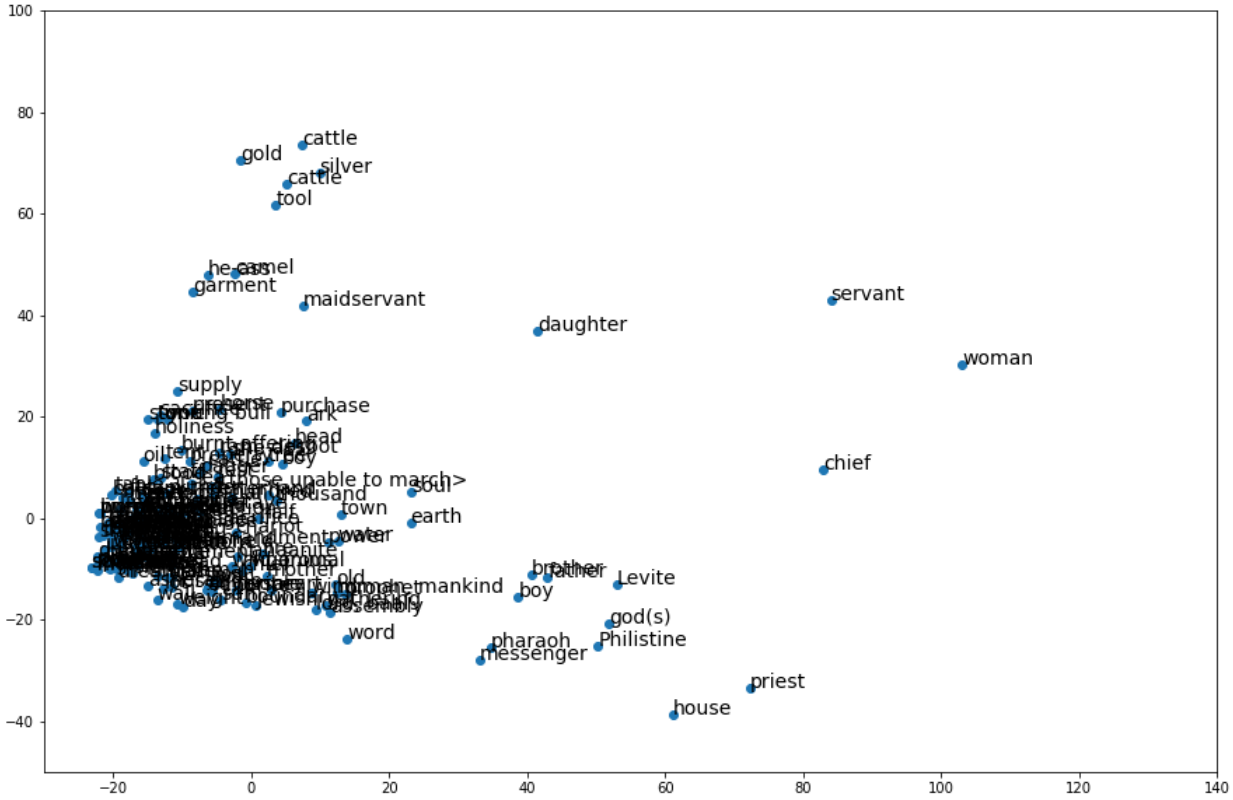
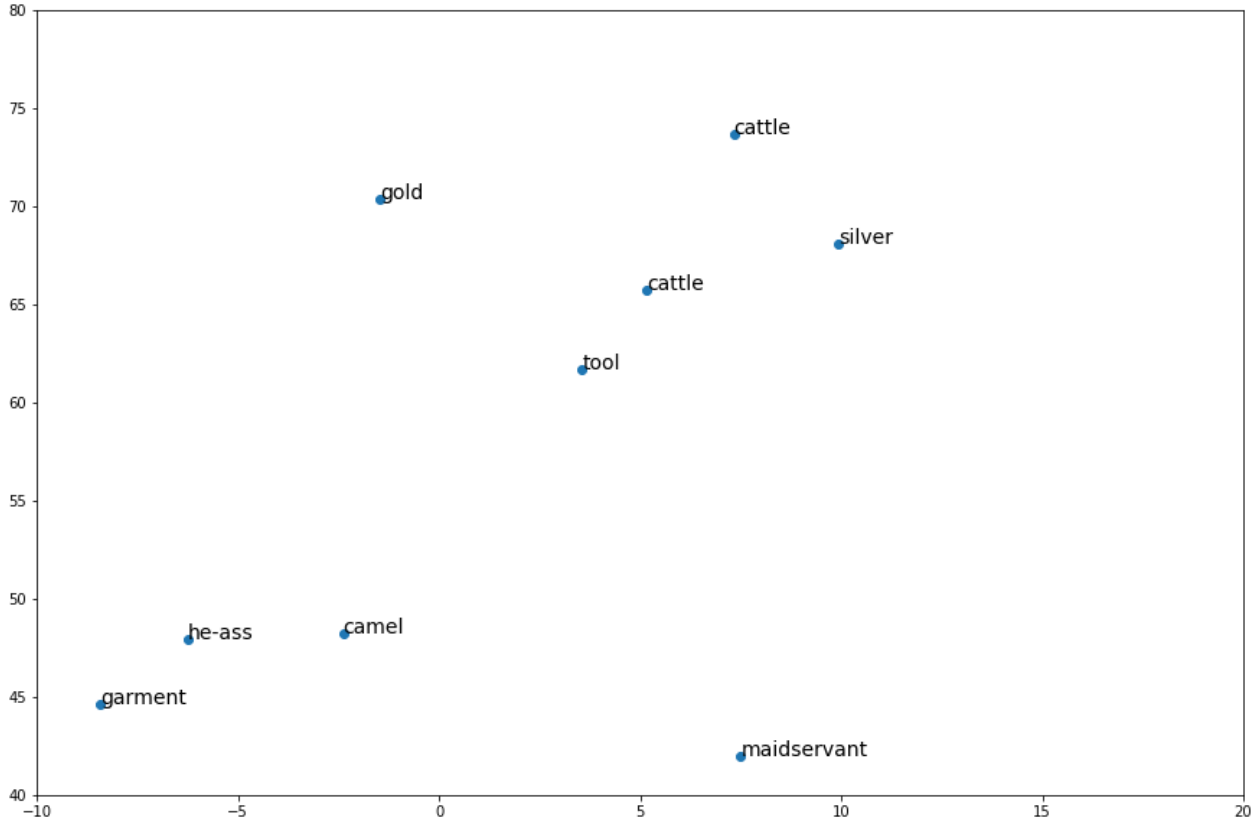
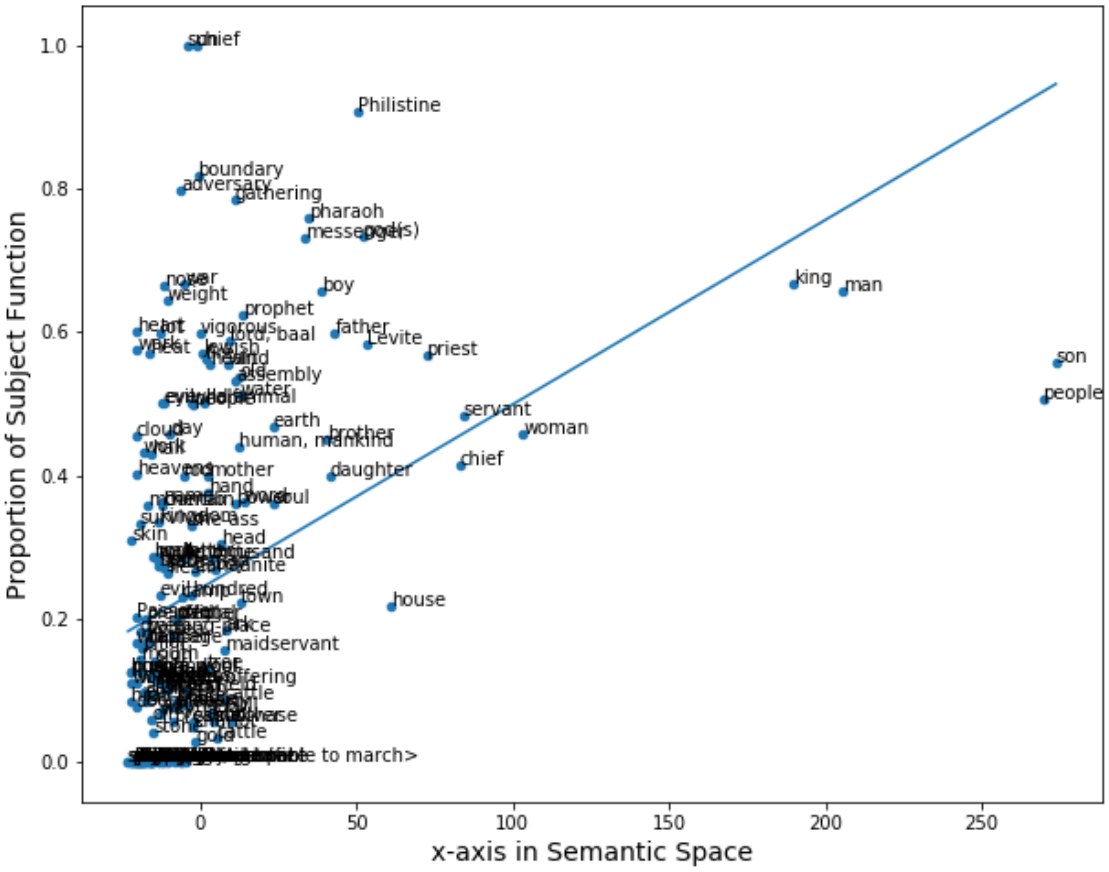


Figure 3. Semantic Space with Log-Likelihood, Possessions Cluster



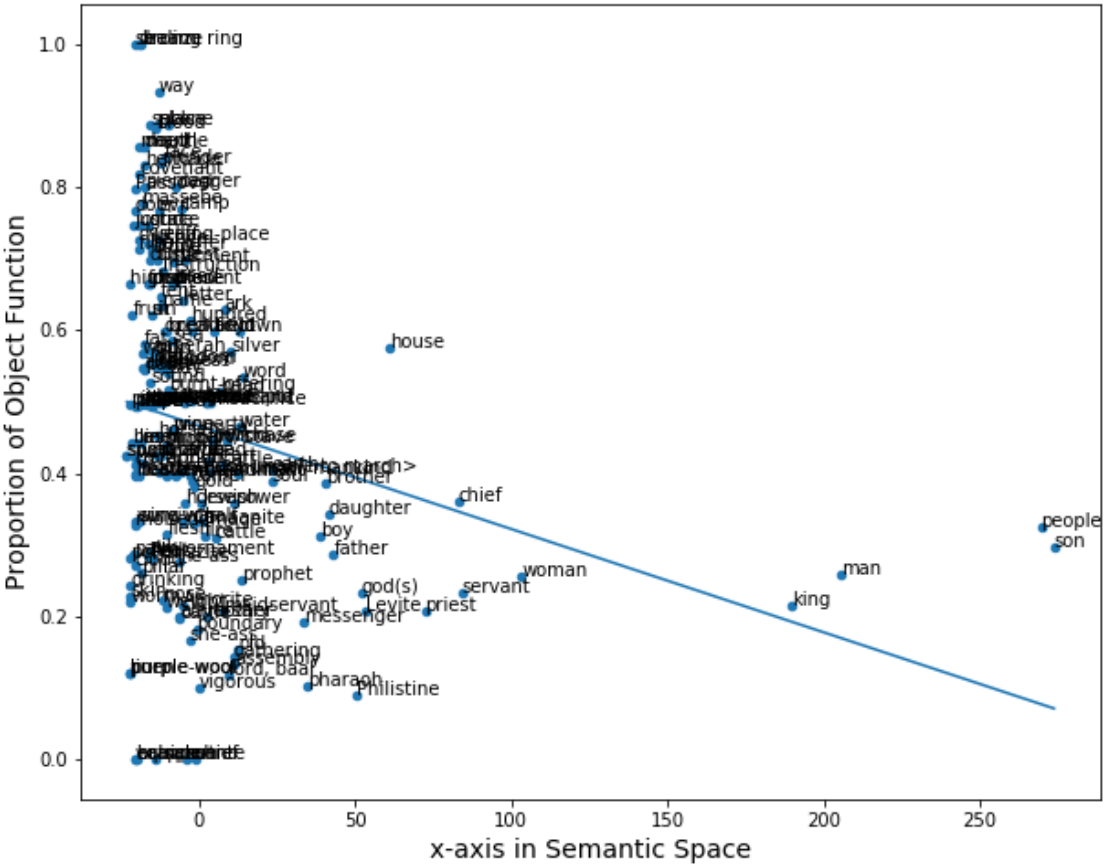
A general pattern can be seen in the plot where people-nouns appear further to the right of the x-axis. To test whether there is indeed a correlation, we can use a linear regression plot. This method is commonly used to test whether two values are correlated. We plot the proportion of times a noun is used as a subject against the values of the semantic space x axis. The results are presented below. Points along the line reflect nouns which follow the hypothesis of correlation.

Figure 4. Proportion of Subject Uses and Distance to Right of Semantic Space



While the points along the line are a bit scattered, likely due to the fact that people-nouns can likewise occur as objects of a verb (and object nouns can appear as subjects), a plot of object relations shows an inverse effect when object phrases are accounted for.

Figure 5. Proportion of Object Uses and Distance to Right of Semantic Space



These opposite trend lines offer at least a preliminary confirmation that nouns that occur more frequently in a subject position cluster further to the right in the semantic space.

Finally, the top similar results per noun show just how promising this approach can be for building noun similarity sets. The results for a few interesting examples are provided below. The target noun is provided above. Beneath it are the top 10 nous rated most similar to it.

BN/ "son"	
>JC/ "man"	0.403
<M/ "people"	0.392
MLK/ "king"	0.372
BT/ "daughter"	0.346
>CH/ "woman"	0.335
PLCTJ/ "Philistine"	0.304
<BD/ "servant"	0.303
FR/ "chief"	0.298
KHN/ "priest"	0.269

DBR/ "word"	
R<H/ "evil"	0.325
TWRH/ "instruction"	0.299
FVN/ "adversary"	0.233
XLWM/ "dream"	0.215
MCPV/ "justice"	0.215
SPR/ "letter"	0.210
XRB/ "dagger"	0.204
MYWH/ "command"	0.202
XQ/ "portion"	0.202

JD/ "hand"	
PNH/ "face"	0.262
KNP/ "wing"	0.240
LB/ "heart"	0.239
>RWN/ "ark"	0.220
MJM/ "water"	0.218
R<H/ "evil"	0.199
MQWM/ "place"	0.189
XRB/ "dagger"	0.181
KP/ "palm"	0.174

KLJ/ "tool"	
KSP/ "silver"	0.332
>WYR/ "supply"	0.304
ZHB/ "gold"	0.292
BFM/ "balsam-tree"	0.281
MNXH/ "present"	0.277
>RWN/ "ark"	0.276
MKWNH/ "place"	0.268
CLXN/ "table"	0.267
KJWR/ "basin"	0.263

<LH/ "burnt-offering"	
XV>T/ "sin"	0.405
XLB=/ "fat"	0.385
MNXH/ "present"	0.360
CLM/ "final offer"	0.288
R</ "evil"	0.285
>JL=/ "ram, despot"	0.271
PSX/ "Passover"	0.254
ZBX/ "sacrifice"	0.252
<PR/ "dust"	0.243

MJM/ "water"	
XKMH/ "wisdom"	0.252
JM/ "sea"	0.221
JD/ "hand"	0.218
JJN/ "wine"	0.210
XJH/ "wild animal"	0.190
KSP/ "silver"	0.179
R<H/ "evil"	0.169
CMC/ "sun"	0.161
<JR/ "town"	0.158

MZBX/ "altar"	
MYBH/ "massebe"	0.479
KJWR/ "basin"	0.356
BMH/ "high place"	0.308
BD/ "linen, part, stave"	0.296
>CRH/ "asherah"	0.272
QVRT/ "smoke"	0.220
MKWNH/ "place"	0.218
KLJ/ "tool"	0.213
MSK/ "covering"	0.20

Full results for the top 50 most common nouns can be perused in the linked repository.[36] As can be seen, a lot of work is still needed, both on the end of the model construction and on the end of post-processing the results. But these early examples show just how valuable the semantic vector space model can be for a form-to-function approach to semantics.

Conclusion

This paper has argued that form-to-function Hebrew linguistics can benefit greatly from developing its methodology for word-level semantics. It has presented the initial results of an experiment which applies semantic vector spaces, using formal structures, to model the general tendencies of nouns in the Hebrew Bible. This method holds great promise for the future of modeling word semantics in the ETCBC database.

References

Abzianidze, Lasha. “A Natural Proof System for Natural Language.” PhD Dissertation, Tilburg Univeristy, 2017. Accessed March 20, 2018. https://pure.uvt.nl/portal/files/14858339/Abzianidze_Natural_20_01_2017.pdf.

de Blois, Reinier. “Towards a New Dictionary of Biblical Hebrew Based on Semantic Domains.” PhD Dissertation, Vrije Universiteit Amsterdam, 2000.

Cook, John A. *Time and the Biblical Hebrew Verb: The Expression of Tense, Aspect, and Modality in Biblical Hebrew*. Linguistic Studies in Ancient West Semitic 7. Winona Lake: Eisenbrauns, 2012.

[36] https://github.com/codykingham/semantics/tree/master/data

Dyk, Janet, Oliver Glanz, and Reinoud Oosting. "Analysing Valence Patterns in Biblical Hebrew: Theoretical Questions and Analytic Frameworks." *JNSL* 40, no. 1 (2014): 1–20.

Firth, J.R. "A Synopsis of Linguistic Theory, 1930–1995." In *Studies in Linguistic Analysis*, 1–32. Oxford: Basil Blackwell, 1962.

Glanz, Oliver, Reinoud Oosting, and Janet Dyk. "Valence Patterns in Biblical Hebrew: Classical Philology and Linguistic Patterns." *JNSL* 41, no. 2 (2015): 31–55.

Goldberg, Adele E. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press, 1995.

Hearst, Marti. "Automatic Acquisition of Hyponyms from Large Text Corpora." *Proceedings of the Fourteenth International Conference on Computational Linguistics* (1992).

Hunston, Susan, and Gill Francis. *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Vol. 4. Studies in Corpus Linguistics. Amsterdam: John Benjamins, 2000.

Kingham, Cody. "ETCBC Data Creation." Last modified March 3, 2018. Accessed March 20, 2018. http://www.etcbc.nl/datacreation/.

———. "Verb in Biblical Hebrew." Data Repository. Last modified 2017. Accessed March 20, 2018. https://github.com/codykingham/Verb_in_Biblical_Hebrew.

Levshina, Natalia. *How to Do Linguistics with R: Data Exploration and Statistical Analysis*. Amsterdam: John Benjamins, 2015.

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

———. *The Role of Zion/Jerusalem in Isaiah 40–55: A Corpus-Linguistic Approach*. Vol. 59. Studia Semitica Neerlandica. Leiden: Brill, 2013.

Padó, Sebastian, and Mirella Lapata. "Dependency-Based Construction of Semantic Space Models." *Computational Linguistics* 33, no. 2 (2007): 161–199.

Roorda, Dirk. "Text Fabric Wiki." Last modified January 27, 2018. Accessed March 20, 2018. https://github.com/Dans-labs/text-fabric/wiki.

Schneider, Wolfgang. *Grammar of Biblical Hebrew*. Translated by Randall Mckinion. Vol. 1. Studies in Biblical Hebrew. New York: Peter Lang, 2016.

———. *Grammatik Des Biblischen Hebräisch : Ein Lehrbuch. Völlig Neue Bearb. Der "Hebräischen Grammatik Für Den Akademischen Unterricht*. München: Claudius, 1974.

Snow, Rion, Daniel Jurafsky, and Andrew Ng. "Learning Syntactic Patterns for Automatic Hypernym Discovery." *Proceeding NIPS'04 Proceedings of the 17th International Conference on Neural Information Processing Systems* (2004): 1297–1304.

Stefanowitsch, Anatol, and Stefan Th. Gries. "Collostructions: Investigating the Interaction of Words and Constructions." *International Journal of Corpus Linguistics* 8, no. 2 (2003): 209–243.

Talstra, Eep. "A Hierarchy of Clauses in Biblical Hebrew Narrative." In *Narrative Syntax and the Hebrew Bible: Papers of the Tilbug Conference 1996*, edited by Ellen Van Wolde, 29:85–118. BibInt. Leiden: Brill, 1997.

———. "An Hierarchically Structured Data Base of Biblical Hebrew Texts. The Relationship of Grammar and Encoding." In *Proceedings of the First International Colloquium, Bible and Computer: Interpretation, Heremeneutics, Expertise. Tübingen, 2-3-4 Septembre 1985*, 337–349. Association Internationale Bible et Informatique. Paris: Champion, 1986.

———. "Exegesis and the Computer Science: Questions for the Text and Questions for the Computer." *BO* 37, no. 3/4 (1980): 121–128.

———. "On Text and Tools. A Short History of the 'Werkgroep Informatica' (1977-1987)." In *Computer Assisted Analysis of Biblical Texts. Papers Read at Teh Workshop on the Occasion of the Tenth Anniversary of the "Werkgroep Informatica," Faculty of Theology, Vrije Universiteit, Amsterdam*, edited by Eep Talstra, 9–28. Amsterdam: Free University Press, 1989.

———. "Text Grammar and Hebrew Bible. I: Elements of a Theory." *BO* 35 (1978): 169–174.

———. "Text Grammar and Hebrew Bible II: Syntax and Semantics." *BO* 34, no. 1/2 (1982): 25–38.

Vossen, Piek. "NewsReader: Using Knowledge Resources in a Cross-Lingual Reading Machine to Generate More Knowledge from Massive Streams of News." *Knowledge-Based Systems* 110 (2016): 60–85.

Xiao, R. "Collocation." In *The Cambridge Handbook of English Corpus Linguistics*, edited by D. Biber and R. Reppen, 106–124. Cambridge: Cambridge University, 2015.