Proceedings of 1st Workshop on Language Technologies for Historical and Ancient Languages, pages 68–73
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

A Thesaurus for Biblical Hebrew

Miriam Azar, Aliza Pahmer, Joshua Waxman
Department of Computer Science
Stern College for Women, Yeshiva University
New York, NY, United States
mtazar@mail.yu.edu, apahmer@mail.yu.edu, joshua.waxman@yu.edu

Abstract

We build a thesaurus for Biblical Hebrew, with connections between roots based on phonetic, semantic, and distributional similarity. To this end, we apply established algorithms to find connections between headwords based on existing lexicons and other digital resources. For semantic similarity, we utilize the cosine-similarity of tf-idf vectors of English gloss text of Hebrew headwords from Ernest Klein's A Comprehensive Etymological Dictionary of the Hebrew Language for Readers of English as well as from Brown-Driver-Brigg's Hebrew Lexicon. For phonetic similarity, we digitize part of Matityahu Clark's Etymological Dictionary of Biblical Hebrew, grouping Hebrew roots into phonemic classes, and establish phonetic relationships between headwords in Klein's Dictionary. For distributional similarity, we consider the cosine similarity of PPMI vectors of Hebrew roots and also, in a somewhat novel approach, apply Word2Vec to a Biblical corpus reduced to its lexemes. The resulting resource is helpful to those trying to understand Biblical Hebrew, and also stands as a good basis for programs trying to process the Biblical text.

Keywords: Corpus (Creation, Annotation, etc.), Less-Resourced/Endangered Languages, Lexicon, Lexical Database, Phonetic Databases, Phonology, Tools, Systems, Applications, graph dictionary, semantic similarity, distributional similarity, Word2Vec

1. Introduction

Biblical Hebrew is the archaic form of Hebrew in which the Hebrew Bible is primarily written. Its syntax and vocabulary differ from later Rabbinic Hebrew and Modern Hebrew. Hebrew is a highly inflected language, and the key to understanding any Hebrew word is to identify and understand its root. For example, the first word in the Bible is בראשית / *bereishit* / 'in the beginning'. The underlying three-letter root is ראש / *rosh* / 'head, start'. By adding vowels and morphology to a root, one can produce derived forms, or lexemes. The lexeme ראשית / *reishit* / 'beginning' is derived from the root ראש. Finally, the prefix letter ב / *be* introduces the preposition 'in'.

Many scholars have developed resources for understanding these Hebrew roots. While we do not intend to provide a comprehensive list, we will mention a few notable resources. *A Hebrew and English Lexicon of the Old Testament*, developed by Brown, Driver and Briggs (1906), is one such standard dictionary. *The Exhaustive Concordance of the Bible*, by Strong (1890), is an index to the English King James Bible, so that one can look up an English word (e.g. "tree") and find each verse in which that word occurs. Strong's Concordance also includes 8674 Hebrew lexemes, and each verse occurrence includes the corresponding Hebrew lexeme number. Some versions of Brown-Driver-Briggs are augmented with these Strong numbers. For example, Sefaria, an open-source library of Jewish texts, includes such an augmented dictionary as part of their database. Another concordance is that of Mandelkern (1896), *Veteris Testamenti Concordantiæ Hebraicae Atque Chaldaicae*, a Hebrew-Latin concordance of the Hebrew and Aramaic words in the Bible, also organized by root.

Another notable dictionary is that of Clark (1999), *Etymological Dictionary of Biblical Hebrew: Based on the Commentaries of Samson Raphael Hirsch*. Rabbi Samson Raphael Hirsch developed a theory, which is expressed through his Biblical commentary (Hirsch, 1867), in which roots which are phonologically similar are also semantically related. This theory is founded on the well-grounded idea, accepted by many scholars, that Hebrew's triliteral roots are often derived from an underlying biliteral root. Thus, the third letter added to the true biliteral root modifies that underlying root's meaning. For instance, Jastrow's dictionary (1903) lists √אב / *`av* is a biliteral root, and derived triliteral roots include אבב / *`avav* / 'to be thick, to be heavy, to press; to surround; to twist; to be warm, glow etc.'; אבד / *`avad* / 'to be pressed, go around in despair', אבר / *`avar* / 'to be bent, pressed, thick', and others. Within Hirsch's system, specific added letters often convey specific connotations.

When comparing roots, alternations between letters within the same or similar place of articulation often carry similar meanings. For instance, in the entry for אבב / *`avav* (listed above), Jastrow notes the connection between it and other biliteral roots, such as קב / *qav*, כב / *kav*, גב / *gav*, חב / *ḥav*, and עב / *'av.* The first letter of אבב, an *aleph*, is a guttural, as is the *ayin* of עב and the *ḥet* of חב. The entry for the triliteral root חבב / *ḥavav,* which is an expansion of the biliteral root חב, includes the gloss to 'embrace (in a fight), to wrestle'. This clearly bears a related meaning to the √אב roots in the previous paragraph, which involved pressing and surrounding. These related meanings might be termed phonemic cognates.

Within the triliteral root system are what might be called gradational variants. At times, there are only two unique letters in a root. For instance, in the root רדד / *radad* / 'flattening down or submitting totally', the two unique letters are the ר / *r* and the ד / *d*. The geminated triliteral root can be formed by gemination of the second letter (as here, the ד / *d* was repeated, to form רדד / *radad*). Alternatively, a hollow triliteral root can be formed by employing a י / *y*, ו / *w*, ה / *h* in one of the three consonant positions. These three letters, *yud*, *vav*, and heh are called *matres lectiones.* They sometimes function in Hebrew as full consonants and sometimes function to indicate the presence of a specific associated vowel. The hollow roots include רדה / *radah* / 'ruling or having dominion over', ירד / *yarad* / 'going down', and רוד / *rod* / 'humbling'. Within Hirsch's system, these gradational variants in general are semantically related to one another, just as is evident in the present case.

While these phenomena have been observed by other scholars, Hirsch made these ideas central to his Biblical commentary and greatly expanded the application of these rules, to analyze many different Hebrew roots. His

commentary on the first verse, and indeed the first word, of Genesis, is typical. In explaining the root ראש / *rosh* / 'head, start' (which has the guttural *aleph* in the middle position), he notes two other words, רעש / *ra'ash* / 'commotion, earthquake' (with a guttural *'ayin* in that position) and רחש / *raḥash* / 'moving, vibrating, whispering' (with a guttural *ḥet* in that position). Hirsch explains that the core phonemic meaning is movement, with ראש / *rosh* being the start of movement, רעש / *ra'ash* as an external movement, and רחש / *raḥash* as an internal movement.

Clark arranged these analyses into a dictionary, and applied the principle in an even more systematic manner. For each headword, he provides a cognate meaning (a generic meaning shared by each specific cognate variant), and discusses all phonemic and gradational variants. In an appendix, he establishes a number of phonemic classes, in which he groups related words which follow a specific phonemic pattern. For instance, he lists phonemic class A54, which is formed by a guttural (א / *aleph,* ה / *heh,* ח / *ḥet,* ע / *ayin*) followed by two instances of the Hebrew letter ר / *resh*. The roots ארר / *`arar*, הרר / *harar*, and ערר / *'arar* mean 'isolate' and חרר / *ḥarar* means 'parch'. These all share a general phonemic cognate meaning of 'isolate'. (To relate the last root, perhaps consider that a desert is a parched, isolated place; perhaps they are not related at all.) A less clear-cut example is A60, which is formed by a guttural, the Hebrew letter ד / *dalet*, and then a sibilant, with a cognate meaning of 'grow'. The roots involved are הדס / hadas / 'grow', חדש / *ḥadash* / 'renew', עדש / *'adash* / 'grow', and עטש / *'atash* / 'sneeze'. There is sometimes a level of subjective interpretation to place these words into their phonemic cognate classes, but some true patterns seem to emerge.

Another noteworthy dictionary is that of Klein (1987), *A Comprehensive Etymological Dictionary of the Hebrew Language for Readers of English*. It focuses not only on Biblical Hebrew, but on Post-Biblical Hebrew, Medieval Hebrew, and Modern Hebrew as well. His concern includes the etymology of all of these Hebrew words, and he therefore includes entries on Biblical Hebrew roots. Klein's dictionary was recently digitized by Sefaria (2018) and made available on their website and their database. Other important digital resources include the Modern Hebrew WordNet project, by Ordan and Wintner (2007), as well as the ETCBC dataset, from Roorda (2015), which provides in-depth linguistic markup for each word in each verse of the Biblical corpus.

Our aim was to create a new digital resource, namely a graph dictionary / thesaurus for the roots (or lexemes) in Biblical Hebrew, in which headwords are nodes and the edges represent phonetic, semantic, and distributional similarity. This captures connections not drawn in earlier efforts. We have thereby created a corpus and tool for Biblical philologists to gain insight into the meaning of Biblical Hebrew roots, and to consider new, possibly unappreciated connections between these roots. The digital resource – a graph database and a Word2Vec model – can also aid in other NLP tasks against the Biblical text – for example, as a thesaurus in order to detect chiastic structures.

2. Method

We sought to create our graph dictionary for Biblical Hebrew in three different ways, creating several different subgraphs. In future work, we plan to merge these subgraphs.

Our first approach was to look for semantic similarities between headwords. Our source data was Ernest Klein's *A Comprehensive Etymological Dictionary of the Hebrew Language for Readers of English*, using Sefaria's (2018) MongoDB database. This dictionary has headwords for both roots (*shorashim*) and derived forms, for Biblical Hebrew as well as many later forms of Hebrew. We first filtered out all but the Biblical roots. Non-root entries have vowel points (called *niqqud*) and non-Biblical Hebrew words are often marked with a specific language code, such as PBH for post-Biblical Hebrew. We calculated the semantic similarity between headwords as the cosine similarity of the tf-idf vectors of the lemmatized words in their English gloss. Thus, אמר / *`amar* and דבר / *dabier* share the English definition 'say', and a cosine similarity of about 0.35. Function words, such as "to" or "an", will have a low tf-idf score in these vectors and would not contribute much to the cosine similarity metric. We therefore set a threshold of 0.33 in creating the "Klein" graph. We applied this approach to Brown-Driver-Briggs' lexicon of lexemes, which had been digitized by Sefaria as well, for the sake of having a comparable graph (for lexemes instead of roots) with semantic relationships calculated in the same manner.

Our second approach was to consider phonetic similarity between headwords. One data source for this was Matityahu Clark's *Etymological Dictionary of Biblical Hebrew*. We digitized a portion of Clark's dictionary, namely his 25-page appendix which contains the listing of phonemic classes containing phonemic cognates with their short glosses. We created a separate graph from this data, linking Clark's headwords to their phonemic class (e.g. ארר to A54) as well as shared short gloss, e.g. ארר / *`arar* to הרר / *harar* based on a shared gloss of 'isolate'.

Aside from that standalone Clark graph, we introduced phonetic relationships on the Klein graph as well. We connected each combination of words which Clark had listed as belonging to the same phonemic class. Additionally, we computed gradational variants for each triliteral root in the Klein dictionary as follows. We treated each triliteral root as a vector of three letters. We checked if the vector matched the pattern of a potential gradational root. If the root contained a potential placeholder letter (י / *yud* in the first position, ו / *vav* or י / *yud* in the middle position, or ה / *heh* in the final position), or if the final letter was a repetition of the middle letter, then it was a potential gradational variant. We then generated all possible gradational variant candidates for this root, and if a candidate also appeared in Klein's dictionary as a headword, we connected the two headwords.

We also looked for simpler, single-edit phonemic connections between headwords in Klein's dictionary. That is, we took the 3-letter vectors for triliteral roots and, in each position, if the letter was a sibilant, we iterated through all Hebrew sibilant letters in that position. We checked whether the resulting word was a headword and, if so, established a phonemic relationship between the word pair. We similarly performed such replacement on other phonetic groups, namely dentals, gutturals, labials and velars.

Our third approach was based on distributional criteria. Our source data was the ETCBC dataset, from Roorda (2015). We first reduced the text of the Bible to its lexemes, using ETCBC lex0 feature. These lexemes were manually produced by human expects. As discussed above, the Hebrew lexeme is often more elaborate than the Hebrew root. Many of the lexemes in this dataset are also triliteral roots (such as ראש / rosh / 'head', and אור / `or / 'light'), but

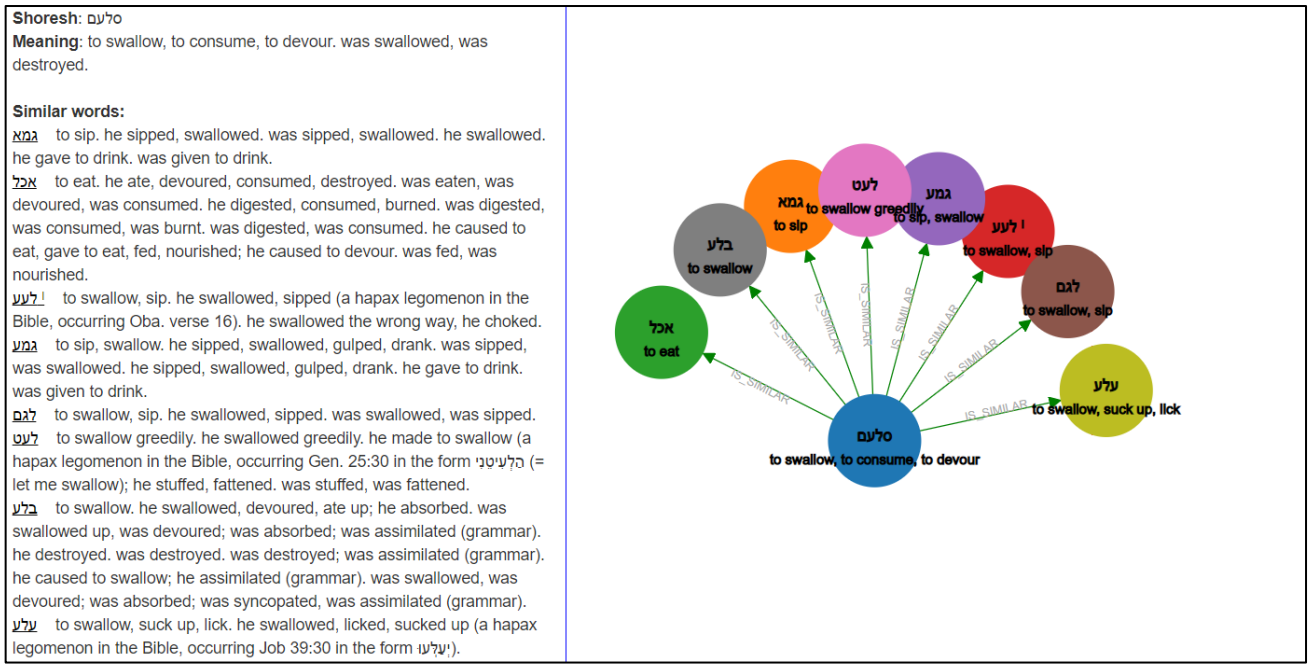

Figure 1: Klein entry for סלעם / *sal'am* / 'to swallow, to consume, to devour'

there are also quite a number of lexemes that would not be considered roots (such as ראשית / *reishit* /'beginning' and מאור / *ma`or* / 'luminary').

We represented each lexeme A as a V-length vector, where V is the vocabulary size (of 6,466). Each position in the vector corresponded to a different lexeme B, and recorded positive pointwise mutual information (PPMI) values. PPMI values of lexeme A and lexeme B were computed as follows:

$$PPMI(A,B) = \max\left(0, log\frac{p(A,B)}{p(A)p(B)}\right)$$

The joint probability p(A, B) is computed as the frequency of lexeme B occurring within a window of the 10 previous and 10 following words of each occurrence of lexeme A, and the individual distributions p(A) and p(B) as the frequencies of lexemes A and B, respectively, within the Biblical corpus.

We then calculated the cosine similarity of each combination of PPMI vectors. Word pairs which exceeded a threshold (again, of 0.33) were considered related. This yielded word pairs such as טוב / *tov* / 'good' and ישר / *yashar* / 'upright' which indeed seem semantically related.

As an additional way of relating words by distributional criteria, we took the same lexeme-based Biblical corpus and trained a Word2Vec model. This is a slightly novel approach to Word2Vec, in that we are looking at the surrounding context of lexemes, rather than the (often highly inflected) full words. The results are promising. For instance, the six most distributionally similar words to ארץ / *`eretz* / 'land' include גוי / *goy* / 'nation', אדמה / *`adamah* / 'earth', and ממלכה / *mamlacha* / 'kingdom', which captures the elemental, geographical, and political connotations of the word 'land'. We filtered by a relatively high threshold of similarity, of 0.9.

We pushed all of these graphs to a Neo4j database and wrote a presentation layer using the D3 JavaScript library. Some of the resulting graphs can be seen at http://www.mivami.org/dictionary, and are also available as a download in GRAPHML file format.

3. Results

By applying our method, we have produced four graphs. Table 1 describes the number of nodes and edges in each graph.

Graph	Nodes	Connections
Klein's Dictionary	3,287 roots	7,472 semantic ; 1,509 phonemic class ; 2,329 phonemic edits
Brown-Driver-Briggs lexicon	8,674 lexemes	12,759 semantic
Clark's Etymological Dictionary	1,926 roots	Grouped into 388 phonemic classes
Distributional Criteria / ETCBC	6,466 lexemes	5773 Word2Vec ; 12,561 PPMI

Table 1: Corpora and Connections Established

At the moment, these different types of connections are in different graphs, and the headword types slightly differ from one another, and so we do not perform a comprehensive inter-graph analysis. However, in the evaluation section, we evaluate the quality of each individual graph, and in this results section, we present some individual interesting subgraphs. We examine the connections between nodes and find that there are some meaningful connections being established.

For instance, Figure 1 depicts the hyperlinked list of related words, from the Klein's dictionary graph, for the root סלעם / *sal'am* / 'to swallow, to consume, to devour'. (In all cases for these graphs, the colors are just the styling provided by the D3 JavaScript visualization library.)

Although the connection to other entries is based on semantic similarities (e.g. sipping, swallowing, gulping), there are some obvious phonological connections between

Shoresh: דבר ׀

Meaning: to speak. (used only in the act. part. דובר, 'saying, speaking', and in the pass. part. דָּבוּר, 'said, spoken'). (pl.) they spoke to one another, talked; was said, was spoken. he spoke of; he spoke to or with. was spoken; was stipulated, was agreed. he spoke, talked; he came to an agreement.

Similar words:

דבב ׀׀ to speak, whisper. he spoke, whispered. he caused to speak; he spoke, whispered. he spoke; he caused to speak, interviewed. he was made to speak.
ברגז to not be on speaking terms with (slang). he was not on speaking terms.
פוץ ׀׀ to speak. he opened his mouth, spoke, said (used only in the perf. and in the part.).
סיח to talk, say. talked, said. he talked, said, spoke.
לאט ׀׀ to speak softly, whisper. he spoke softly, whispered (a hapax legomenon in the Bible, occurring Job 15:11).
שיח to speak, talk, converse. he spoke, talked, conversed; he mused, meditated. he talked. , he spoke, talked conversed; tr. v. he caused to speak, caused to talk.
לוז ׀׀ to speak evil, slander. he spoke evil of, slandered.
נום ׀׀ to speak. he spoke, said.
חטם ׀׀ to speak through the nose. he spoke through the nose. (of s.m.).
מלל ׀ to speak, say, utter. he spoke, said, uttered. was spoken, was said, was uttered. he spoke with.

Figure 2: Klein hyperlinked entry for דבר

these roots. In particularly, the letters לע / *lamed-'ayin* appear in many words, as well as גמ / *gimel-mem* and לג / *lamed-gimel*. Sounding out each of these words, they all feel quite onomatopoetic, imitative of the sound of sipping and swallowing.

The connections in the Klein graph can, more generally, function as a thesaurus, providing insight into the inventory of similar words conveying a concept. Someone using Klein's print dictionary could look up the word דבר / *dabeir*, and discover that it means 'speak'. However, what similar words could the Biblical author have employed? Figure 2 shows the hyperlinked list of 'speak' words:

Interestingly, the common word אמר/ *`amar* / 'say' does not appear in this list, because 'say' did not appear in the entry for דבר, only 'speak'. It is, however, in the two-step neighborhood of דבר, because it is a neighbor of the root מלל / *maleil* / 'to speak, say, utter'.

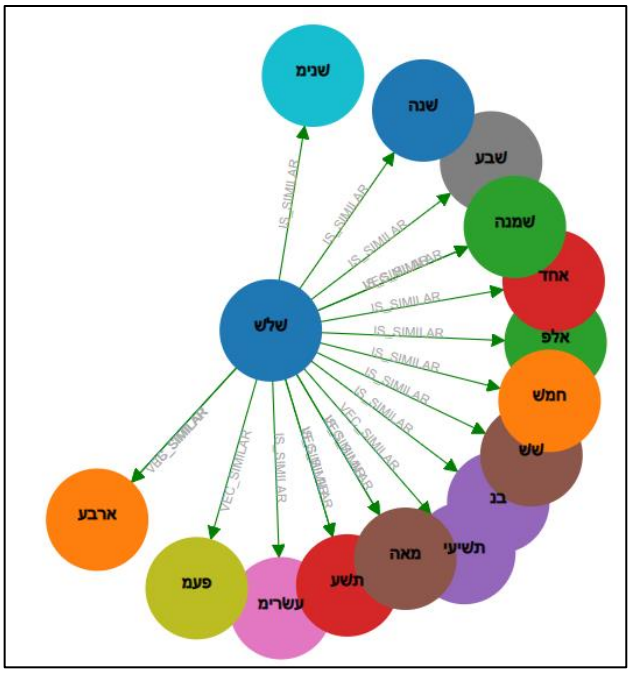

Figure 3: Distributional entry for the word שלש / *shalosh*

Meanwhile, an examination of sample entries in the distributional graph reveals real connections between words. For instance, Figure 3 displays the graph for the word שלש / *shalosh* / 'three'. The connected entries are for many other numbers, such as אחד / *`eḥad* / 'one', שבע / *sheva'* / 'seven', and אלף / *`eleph* / 'thousand', as well as the word פעם / *pa'am* / 'occurrence' and שנה / *shanah* / 'year'. Some of these connections are based on Word2Vec, some on PPMI vector similarities, and some on both.

Finally, the present version of the Clark graph simply shows roots linked to their phonemic classes, as well as connections between roots whose short translation is identical. Since the connections are essentially manually crafted, the graph is exactly as we would expect. Figure 4 shows the graph for the Clark entry of המר / *hamar* / 'heap'.

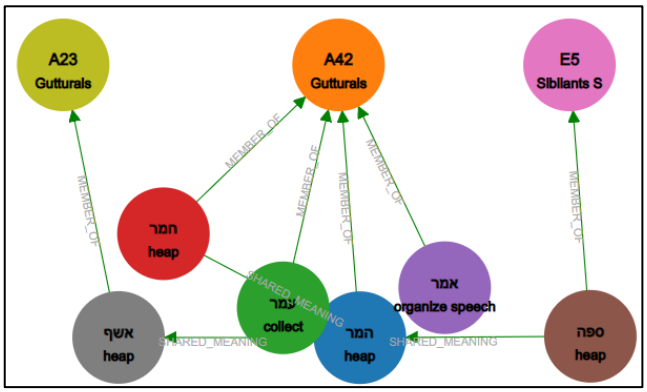

Figure 4: Clark entry for המר / *hamar*

If we had examined the same entry המר / *hamar* in Klein's dictionary, the gloss would be 'to bet, enter a wager'. This might be an example where Clark's decision as to the proper definition of המר / *hamar* was influenced by a desire to structure all A42 phonemic cognates into related words. When interpreting a specific instance of the word, one would need to carefully consider the Biblical usage, in context.

Consider how אמר / *`amar*, usually rendered as 'say', here is explained as 'organized speech', so that it works well with other roots which mean 'heap' and 'collect'. This root is placed in the phonemic class A42, which appears to be formed by a guttural as the first letter, followed by מ / *mem* and ר / *resh*. The subgraph also shows other roots, from other phonemic classes, with a shared meaning (namely "heap"), along with the phonemic class of those roots. This is a fitting way of exploring words within the context of their phonemic cognates.

4. Evaluation

To evaluate the precision of the semantic connections that we discovered within the Klein dictionary, we outputted and analyzed all connections between headwords that exceeded our 0.33 threshold of cosine similarity.

Among the 3287 Klein dictionary roots, 2728 were connected to another root, and we established 7472 such semantic relationships, for an average of 2.73 connections per word. However, a closer examination of the graphs reveals a number of tightly connected subgraphs or even cliques. That is, the graph contains several subgraphs in which a large number of semantically related roots link to each another. For instance, אגד / *`agad* contains a number of

word senses including 'to bind, tie, join together, unite, amalgamate'. It is connected to 15 other roots, including אחד / *`eḥad* / 'to be one, to unite' – also a phonetically related word, אסר / *`asar* / 'to bind, tie, imprison', and חבר / *ḥabar* / 'to be united, be joined'. The first listed word אחד / *`eḥad*, is connected to 4 other words, 3 of which have the 'join' meaning. The word אסר / *`asar* is connected to 10 other words, all of which have the 'tie / bind' meaning. And the last word, חבר / *ḥabar* is connected to 8 words, which all have the 'join / attach' meaning.

We submitted the Klein root connections to human experts for judgement, to determine if the headwords indeed had semantic similarity. Of the 7472 connections, 6793 were deemed correct, for a precision of 0.91. We examined the 9% mistaken connections and found that the vast majority (539 out of 679, or 79%) were the result of three filtering errors particular to our dataset. Namely, often the gloss for a root was simply that this was a "base" (that is, a root) for a different non-root headword, that we should "see" the definition in another headword, or that the word was a "hapax legomenon", that is, a word which occurs only once in the Biblical corpus and can therefore only be guessed at based on context. The vectors for these glosses were similar, but not based on real semantic content. A fix would entail filtering out such null-glossed words, and linking the cross-references.

Most of the remaining erroneous connections were due to homonyms and homographs within the stemmed English gloss words. For instance, "tear" can either be a droplet from the eye or the act of ripping something, "left" can either be the opposite of right or the act of going away, and "leave" might refer to the act of going away or to a tree leaf. A few errors were due to non-essential function words, e.g. "to cut *off*" and "to hollow *out*". A fix might entail including part of speech disambiguation in the vectors, or comparison with a similar dictionary in another language.

We performed similar analysis among the lexemes in the Brown-Driver-Briggs lexicon, and found similar results to our results for Klein's Dictionary. Among the 8674 lexemes, 5047 were connected to another lexeme. We established 12,760 semantic relationships, for an average of 2.52 connections per word. We subjected 500 of these relationships to human judgement, which yielded a precision of 0.76. Among the correctly discovered relationships, we discovered some tightly connected subgraphs.

We analyzed the errors and could not find ready explanations for the vast majority of them. The corpus is quite different from Klein's dictionary. While Klein has headwords as roots, with what might be considered lexemes grouped together into a single entry, Brown-Driver-Briggs separates these lexemes into different entries. Each entry includes fewer words and English synonyms. Brown-Driver-Briggs also contains entries for Biblical personages, with a discussion of the etymology of their name, plus that they were the son, daughter, father, or mother of some other person. This has the effect of linking etymologies with familial relationships, and unrelated etymologies together by way of the familial relationships – that is, it introduces a good deal of noise. A fix would entail filtering out these Biblical names, but perhaps vector similarity is not as suited for shorter gloss entries.

We performed a similar analysis for the PPMI vector-based distributional approach applied to lexemes from the ETCBC dataset, where the threshold of cosine similarity of the vectors was 0.33. Of the 6466 lexemes, 4478 were connected to another lexeme. We established 12,561 connections, for an average of 2.80 connections per lexeme. A sample of 200 connections were reviewed by a human expert, where any relationship between the two lexemes (and not just synonymy) was deemed correct. The precision was 0.82. The majority of relationships found (67%) were between names of people or places, appearing for instance in Biblical genealogical lists or descriptions of borders, since these names occur rarely and only in context of each other. There were meaningful connections found. For instance, עדשה / *`adašaha* / 'lentil' is mentioned in II Samuel 17:28 among other places, and connections are made to the other grains and foodstuffs listed in the verse, but not to the beds, basins, and earthen vessels.

We similarly evaluated our Word2Vec approach. We set a relatively high similarity threshold of 0.9. We connected 1209 lexemes to one another, establishing 5772 connections, or about 4.8 connections per lexeme. Human evaluation of 200 such connections yielded a precision of 0.98. While a majority were again person and place names, those which were not were highly related to one another, e.g. the ordinal numbers, antonyms such as "light" and "darkness", and synonyms such as types of grass. As is typical of Word2Vec, by lowering the threshold, we encounter more connections which are more tangential but still related. In general, for all of these graphs, further exploration is needed regarding where to set the threshold parameter.

Our assessment of the precision of phonetic relationships on the Klein graph was performed programmatically, by checking whether the semantic similarity of the tf-idf vectors exceeded the 0.33 threshold. Table 2 shows the precision for each type of connection.

Connection Type	# Connections	Precision
Cognate Class	1509	0.03
Gradational Variant	275	0.11
Guttural replacement	582	0.07
Velar replacement	208	0.02
Sibilant replacement	168	0.24
Labial replacement	398	0.02
Dental replacement	698	0.01

Table 2: Connections for Phonological Relationships

Certain phonetic relationships – most notably sibilant replacement at 24% and gradational variants at 11% – seem to be borne out and valuable. Other relationships – such as dental replacement and belonging to the same phonemic class defined by Clark, do not seem to be borne out.

This might demonstrate that these phonetic connections and phonemic classes were an overreach, the result of trying to globally impose a system that works between certain word pairs but does not hold in the general case. Alternatively, the theory of phonemic classes – that there is a basic cognate meaning, with individual letter choices modifying this basic meaning in particular directions – involves a different approach to describing the word's meaning, one which is not captured by an English gloss which does not carry such concerns. For instance, עדש / *'adash* is the root of lentil (as above), which is something that grows. Clark connects it to other growing / renewal words, but he would not expect Klein to mention growing, rather than lentils, in his gloss. Similarly, Hirsch would not be at all surprised that a standard

dictionary would not relate ראש / *rosh* / 'head' to רעש / *ra`ash* / 'earthquake' and רחש / *raḥash* / 'vibrate'. Perhaps some of these relationships could be reproduced by considering a lower semantic similarity threshold, by considering Word2Vec distributional similarity, or by a WordNet ontology, but perhaps not.

Additionally, we would note that the low precision in some types of transformation simply indicates that while phonetically related words might be semantically related, this is not necessarily systematic, for all possible combination of gutturals (or velars, etc.) and for all letter positions. Additional exploration of the phonetic transformations with the greatest semantic value is necessary.

5. Future Work

We would like to develop a heuristic to stem the lex0 features in the ETCBC dataset to be roots rather than lexemes, so as to consider distributional criteria of roots, as well as to be able to create these connections on the Klein graph, which works with roots. We would like to similarly reduce entries in the Brown-Driver-Briggs lexicon to such roots, again to create a unified graph to enable a valid, apples-to-apples, quantitative evaluation.

With all these connections in place, we hope to apply machine learning, to discover which types of letter substitutions are likely to yield related terms, and to give a measure of the phonemic relatedness of two root entries.

Also, at the moment, within semantic similarities, we are primarily finding synonyms. We would like to expand the types of connections between entries, to find antonyms and hypernyms. There has been some recent work on finding such relationships using Word2Vec vectors, and so we could find such relationships based on our distributional graph. For the semantic similarity graphs, we could harness an English resource such as WordNet applied to the English gloss text of the Klein entries.

There are a few Digital Humanities projects that we look forward to implementing using the corpus as it presently stands. One such project involves detection of chiastic structure in the Biblical text, and the parallel words we need to detect are often synonyms rather than exact repetition of the root. Finally, we would look to duplicate this thesaurus construction process for other Semitic languages, such as Arabic or Aramaic and consider cross-lingual connections.

6. Acknowledgments

We would like to thank the Drs. Phyllis & Abraham S. Weissman Summer Student Research Fund in STEM at Stern College for Women for providing the initial funding for the research, from which the present project developed.

7. Bibliographical References

Brown, F., Driver, S.M., Briggs, C.A. (1906). A Hebrew and English Lexicon of the Old Testament, England.

Clark, M. (1999). Etymological Dictionary of Biblical Hebrew: Based on the Commentaries of Samson Raphael Hirsch. Feldheim, Nanuet, NY.

Hirsch, S. (1867 – 1878). Uebersetzung und Erklärung des Pentateuchs, 5 vols. Frankfurt-on-Main, Germany.

Klein, E. (1987). A Comprehensive Etymological Dictionary of the Hebrew Language for Readers of English. Carta, Jerusalem, Israel.

Mandelkern, S. (1896). Veteris Testamenti Concordantiæ Hebraicae Atque Chaldaicae. Veit et Comp., New York

Ordan, N., Wintner, S. (2007). Hebrew WordNet: a test case of aligning lexical databases across languages. International Journal of Translation 19(1):39-58, 2007.

Roorda, D. (2015). The Hebrew Bible as Data: Laboratory - Sharing – Experiences, https://arxiv.org/pdf/1501.01866.pdf

Sefaria, 2018. New Releases: Jastrow and Klein Dictionaries https://blog.sefaria.org/blog/2018/11/12/jastrow-and-klein-dictionaries

Strong, J. (1894). The Exhaustive Concordance of the Bible. Hunt and Eaton, New York.

8. Language Resource References

Eep Talstra Centre for Bible and Computer. (2015). ETCBC Dataset

Sefaria (2018). The Sefaria MongoDB.