Index structures in early Hebrew Biblical word lists

Preludes to the first Latin concordances

Bella Hass Weinberg

The earliest Hebrew Masoretic Bibles and word lists are analyzed from the perspective of index structure. Masoretic Bibles and word lists may have served as models for the first complete Biblical concordances, which were produced in France, in the Latin language, in the 13th century. The thematic Hebrew Biblical word lists compiled by the Masoretes several centuries earlier contain concordance-like structures – words arranged alphabetically, juxtaposed with the Biblical phrases in which they occur. The Hebrew lists lack numeric locators, but the locations of the phrases in the Bible would have been familiar to learned people. The indexing methods of the Masoretes are not known, but their products contain many structures commonly thought to date from the modern era of information systems, among them word frequency counts, distinction of homographs, positional indexing, truncation, adjacency, and permuted indexes. It is documented that Hebrew Bibles were consulted by the Latin concorders; since Masoretic Bibles had the most accurate text, they were probably the editions consulted. This suggests the likely influence of Masoretic lists on the Latin concorders.

Introduction

This is the third in a series of papers on the history of indexing prepared for this journal. The first one (Weinberg, 1999), based on research at the Vatican Library, established that indexes originated in the domain of religion, while the second (Weinberg, 2000a) described the earliest indexes compiled in France, where subject indexing was allegedly invented (Rouse and Rouse, 1979: 11). The current paper reports on the findings of investigations that were suggested as ‘Directions for further research’ in Weinberg (2000a: 10–11). Some of the results were reported in a paper presented at the Conference of the International Federation of Library Associations (IFLA) held in Jerusalem in August 2000, of which a brief electronic preprint is available (Weinberg, 2000b).

The previous Indexer paper described the first Biblical concordances in the Latin language and suggested that Hebrew Biblical word lists compiled by the Masoretes may have served as the models for them. Research on this subject, conducted in Israel in the summer of 2000, included examination of the earliest Masoretic Bibles and word lists, of which microfilms are housed in the Jewish National and University Library in Jerusalem, as well as interviews with experts on the Mesorah. In July 2001, the original manuscripts of early Masoretic Bibles in the British Library and Oxford University Library were examined.

This article assumes no knowledge of Hebrew or Latin on the part of the reader, but it does assume familiarity with basic indexing terminology and index structure, though information science terms are defined. At the outset, let us review some basic facts about the Bible. The Old Testament was originally written in Hebrew. The New Testament is not part of Jewish Bibles; thus the term Bible in a Jewish context, as in the English title of Even-Shoshan’s Hebrew concordance (1977–80), includes only the three parts of the Old Testament – Pentateuch, Prophets, Writings – for a total of 24 books.

What is the Mesorah and who were the Masoretes?

The Hebrew word mesorah (synonym masoret)1 means ‘tradition’. The purpose of the ba’ale mesorah (‘masters of the Mesorah’) or Masoretes was to preserve the Hebrew Biblical text and transmit it accurately to subsequent generations. (The Latin concorders had a similar purpose, as was pointed out in Weinberg (2000a: 10).) The transmission was initially done orally; later it was written down. The Masoretes continued the work of the Sofrim, generally translated as ‘scribes’, but literally ‘those who count’. The Sofrim counted the letters in each book of the Bible to ensure that the text would neither be added to nor deleted from. (Blau (1896) demonstrated that there were errors in the totals.) The Masoretes incorporated the work of the Sofrim (see Fig. 1) and, in addition, noted variant spellings and created a variety of word lists, some of which are indexes.

Standardization of the text of the Hebrew Bible began in the 4th and 5th centuries CE (Even-Shoshan, 1977: [xvii]), but notes could not be written in the parchment scrolls used for public reading. After the adoption of the codex (manuscript book) by Jews in the 6th or 7th century (Dotan, 1971: col. 1409), Masoretic notes were written between columns or in the margins of the Hebrew Bible. The brief, generally abbreviated, notes written between columns are called the
Small Masorah; the longer notes, generally written in the top and bottom margins, and occasionally on the side, are called the Great Masorah (Dotan, 1971: cols 1419, 1424). An important manuscript containing both types of Masoretic notes is held by the British Library (see Figs 1–4). This codex is significant because it mentions the famous Masorete Aaron Ben-Asher and hence can be dated to the 10th century. (This codex is cited henceforth as Or. 4445; i.e. Oriental manuscript 4445).

Around the 9th century, separate word lists were compiled. These were sometimes written at the end of Biblical manuscripts; hence the term Final Masorah (Dotan, 1971: col. 1419). In other cases, the lists constituted separate works and are categorized as the Independent Masorah (Dotan, 1971: col. 1428). The best-known compilation of Hebrew Biblical word lists without accompanying Biblical text is Okhlah ve-okhlah, which is believed to have been edited in the 10th century, although it has ancient elements (Dotan, 1971: col. 1428). The title of the work is derived from the initial words in the first list, which juxtaposes words with and without the particle ve- ('and'). This Hebrew particle is attached to the word. The Masoretic list records the Biblical phrases in which the variants are found. Important manuscripts of Okhlah ve-okhlah are held by the Bibliothèque Nationale in Paris and the library of the University of Halle (Germany).2 The lists in these manuscripts look like indexes, with words and phrases aligned in columns.

Indexing structures in Masoretic lists

That the Masoretes were the predecessors of concordance makers has already been pointed out by at least two authors. Even-Shoshan (1977: [xvi]) noted this in the introduction to his concordance to the Hebrew Bible, and Wellisch (1985: 72) stated this in his article on Hebrew concordances. I have encountered two uses of the term ‘Masoretic concordance’: in an encyclopedia article (Levias, 1904: 365) and in the subject index to a manuscripts catalog (Davis, 1978: 381). To my knowledge, however, no one has to date done a detailed analysis of Masoretic lists from the perspective of index structure, and that is the purpose of this article.

Publications on the Mesorah are issued primarily in the domains of Biblical and Jewish studies, not information science. These publications focus on a variety of facets of the study of the Masorah.
Mesorah, including Hebrew vowel points (Hebrew is a consonantal script) and cantillation marks (musical notes for public readings of the Bible). Dotan's (1971) encyclopedia article on the Mesorah devotes a great deal of space to these topics. This paper excludes those facets and is limited to the lexical (word) and morphological (word-forming element) levels of the Hebrew language.

In English, the form of nouns and verbs hardly varies in different syntactic cases, but in Hebrew, nouns are declined (though not as extensively as in Latin) and verbs conjugated. The Masoretes created lists of all the morphological variants of a root (most Hebrew words have a three-letter root) found in the Bible. Because such lists are not arranged alphabetically, they are of greater interest to grammarians than indexers. Other morphological lists created by the Masoretes are germane to this paper and are illustrated below.

Many Masoretic word lists are arranged in the order of the Hebrew alphabet, a known order of symbols, which is essential for index arrangement. Most Masoretic lists have a title indicating their scope; where the list is arranged alphabetically, this is indicated in the title through the Hebrew word *alefbet* (alphabet). Not all the alphabetical lists are indexes, however. For example, leaf 100b of Or. 4445 records the unique two-letter words in the Bible (see Fig. 2). No location information is given; this is thus not an index.

Many Masoretic lists consist of two elements: a word followed by the Biblical phrases in which it occurs. The word serves as the *heading* of the index entry. In Latin concordances, the Biblical phrase following the heading serves as the *modification*, and the number of the chapter as the *locator*. A millennium-and-a-half before chapter numbering was introduced by Stephen Langton (in the 13th century), the text of the Hebrew Bible was broken up into sections for the purpose of public readings (see Fig. 1). These sections, called *sedarim*, were sometimes numbered in Hebrew letters, but the numbers were rarely used in references. In the absence of numbering of sections of the text in most early Hebrew Bibles, the phrase may be considered the locator in Masoretic lists. The phrase provides enough information for the knowledgeable reader to recognize its location in the 24 books of the Hebrew Bible.

Where there are multiple occurrences of a word in the Bible, the Hebrew phrases are arranged in canonical order; that is, the order of the books of the Bible. Early Latin concordances also arranged the phrases canonically, as the examples in Weinberg (2000a: 5) demonstrate. This is

---

**Figure 2.** The right-hand margin contains a list, in alphabetical order, of unique two-letter words in the Bible, concluding with a title beginning with the Hebrew word *alefbet* (alphabet). Leaf 100b of Mus. Or. Ms 4445. By permission of the British Library.
The scope and variety of Masoretic lists

The Masoretes did not compile a general concordance to the Hebrew Bible; they compiled thematic lists. These are remarkable both in terms of their number and variety. Okhlah ve-okhlah contains close to 400 lists.

Lists of words with and without the attached conjunction ‘and’ in Okhlah ve-okhlah have already been mentioned. In Hebrew, some prepositions are also attached to the following word, and Masoretic lists compare these. For example, there are alphabetical lists of words beginning with the prefixes ke- (‘like’) and be- (‘in’). The Hebrew letters bet and kaf are similar except for a horizontal stroke on the bet. The lists served to ensure that the appropriate letter was written in Biblical manuscripts. Dotan (1971: col. 1421) explains that many notes in the Small Masorah ‘point out forms in the text concerning which there is some apprehension that the reader or scribe-copyist will err’.

This discussion of conjunctions and prepositions leads to the subject of function words – words that serve a grammatical function only, which are omitted from many concordances and word indexes. The first Latin concordances excluded them; the function words in the Latin Bible were indexed in the 15th century because of their importance in theological debates (Walker, 1894: 9). Many centuries earlier, however, the Masoretes compiled lists of verses containing multiple function words, such as et, et, ve-et,5 no doubt to ensure that these little words would not be omitted by scribes.

In computer-generated word indexes, function words are often placed on stoplists (lists of words not to be indexed) because of their frequency. The frequency of words was studied in the pre-computer era, notably by George Kingsley Zipf (1949), who found that function words are the most frequent words in any text. Zipf’s law was applied to automatic indexing, as discussed in Weinberg (1981). Imagine my delight in discovering that Masoretic lists included frequency counts a millennium earlier. A typical entry gives the word, a statistic, and the Biblical phrases in which the word is found. Frequency counts, or postings data (hits), are the primary feedback mechanism in on-line searching (Weinberg and Cunningham, 1983). They also inform the relevance-ranking algorithms of Internet search engines. In most automatic indexing algorithms, words of frequency one are considered insignificant (Resnikoff and Dolby, 1972: 173–4), but the Masoretes highlighted such words, marking them with a code meaning ‘there is no other occurrence of this form in the Bible’.

Rouse and Rouse (1979: 9–11) state that indexes were invented to assist preachers in developing sermons. While the Masoretes had a different purpose, the frequency counts combined with the phrases were surely helpful to writers of Biblical commentaries. Masoretic lists are in fact cited by Jewish commentators, the most famous of whom is Rashi, who lived from 1040 to 1105, in France.

According to Rouse and Rouse (1979: 8), the predecessors of concordances were distinctions, 12th-century tools that distinguished the multiple meanings of a word in the Latin Bible. The Masoretes did not operate only on the graphic level of language; they considered the semantic level as well and distinguished the meanings of words in the Hebrew Bible that have identical spelling. Similarly, a major activity in thesaurus construction is distinguishing the meanings of homographs, words spelled alike but with different meanings.

The Masoretes counted the occurrence of common words and phrases at the beginning or end of a Biblical verse. For example, leaf 60a of Or. 4445 states: ‘atáh (‘now’) 25 times at the head of a verse (see Fig. 3); leaf 105b reports that the phrase ‘i am the Lord your g-d’ occurs 24 times at the end of a verse. This is analogous to positional indexing in full-text databases, whereby the computer records the position of each word in a field or sentence. With sophisticated interfaces, one can request searches on the beginning or end of a text string. A simple example is searching the initial words of a book’s title.

Closely related to this structure is Masoretic indexing of two-word phrases. Leaf 100a of Or. 4445 indicates that the phrase ‘al mayim (‘on water’) occurs five times in the Bible (see Fig. 4). A Hebrew Bible held in Oxford6 lists the occurrences of the phrase tel ‘olam (‘an everlasting ruin’) (see Fig. 5). This is comparable to the adjacency feature of modern search systems – requesting that two words occur next to each other in the specified order. (A simple keyword search on the words ‘information’ and ‘science’ will retrieve documents on ‘information science’ as well as ‘science information’, but an adjacency search on one of these expressions will exclude the other.) Information scientists have attempted to develop algorithms for computer identification of significant two-word expressions, especially compound nouns, for automatic indexing. Warner (1994: 271–2) surveys the computational linguistic issues. The Masoretes identified significant collocations using human intelligence alone.

Leaf 96b of Or. 4445 indicates that 11 verses of the Bible begin with the Hebrew letter nun and conclude with the final form of the same letter. Davis (1978: 219) describes a Masoretic manuscript7 that lists verses ending with double nun, the medial and final forms of the letter. Such indexing may be viewed as the predecessor of truncation, searching on the initial or final letters of a word. A common English-language example of the latter is a search on words ending in -mycin, to retrieve a class of antibiotics. Although the Hebrew and English examples have different applications, they are structurally similar.

Permutred indexes, indexes that ‘flip’ the order of words, often thought to have been invented in the computer era, are also found in Masoretic lists. For example, one list in the Halle manuscript contrasts occurrences of the phrase mishpat tsedek (‘[a] judgement of righteousness’) with those of the expression tsedek u-mishpat (‘righteousness and justice’). Another notes the occurrences of the phrases ‘God said to Moses’ and ‘Moses said to God’. The largest permutred index published today is the Permuterm Index, a section of the citation indexes issued by the Institute for Scientific Information (Garfield (1976) describes its structure). Here we contrast this contemporary permutred index with the related lists of the Masoretes: the latter record comparable to the arrangement of subheadings in a book index in page number order.
actual permutations found in the Bible, while the Permuterm Index, which is generated by computer, creates all possible combinations of significant terms in article titles (Garfield, 1976: 289).

In sum, Masoretic lists contain many index structures thought to have been invented in the 20th century. Hahn (1998) has shown that many indexing techniques and search capabilities of Internet search engines originated several decades earlier in the era of on-line searching. Here we have demonstrated that some of those features are more than a millennium old.

Graphic and artistic considerations

The layout of Masoretic Bibles has already been mentioned: brief notes on orthography or frequency were often written between columns; alphabetic lists were usually placed in the outer margins, and sometimes in the top or bottom margins. The position of Masoretic notes on a page is germane to an analysis of index structure. The marginal notes and lists generally relate to a word contained in the text on the same page. The list of occurrences of the same word elsewhere in the Bible may be viewed as see also references or hyperlinks to related passages.

The Masoretic lists in the margins of Biblical texts may be compared with the embedded citation indexes that are described in Weinberg (1997: 318–20). Those indexes lead from a passage in the Talmud to subsequent legal codes that cite it. The Talmud also includes embedded references to parallel passages. Marginal Masoretic lists are analogous to those in that they refer the reader to other parts of the Bible where the same word occurs.

The difference between Talmudic citation indexes and marginal Masoretic lists is that the appropriate arrangement of the former is canonical (the order of the text), while the placement of a Masoretic list within the text requires the
user to know at least one place in the Bible where the word of interest occurs – if the scribe repeated the list on all the relevant pages, which was not always the case (S. Leiman, personal communication). A full alphabetical concordance makes fewer demands on the user. The difference is comparable to multiple indexes vs a single, merged index. The former require users to think of the appropriate category for their search arguments (name, title, subject) while the latter requires only knowledge of the alphabet.

Some Masoretic lists were written in micrography, a unique Hebrew art form in which tiny letters form a design. The beauty and intricacy of this writing style inspire awe, but Beit-Arié (1993: 86) states that when Masoretic notes were contained in elaborate images, ‘the secondary text lost its verbal meaning altogether and was transformed into a purely visually expressive tool’. Dotan (1971: col. 1427) observes that such ornamentation was provided by copyists who did not understand Masoretic notes; he considers these manuscripts ‘worthless for the study of the Masorah’. Thus Masoretic notes in ‘plain text’ are of far greater interest to indexers than the artistic presentations, although the latter are often reproduced in books of ‘treasures’ of libraries.

Theories of the Masoretes’ knowledge and working methods

The lack of section numbering in early Hebrew Bibles leads to the question: how did the Masoretes compile their indexes without locators? Did they know the entire Bible by heart? Could they recall all occurrences of a function word from memory?

The scholars I have consulted are divided on the issue of whether the Masoretes memorized the 24 books of the Old Testament or had an indexing system to facilitate their work.
Masoretic lists are not error-free, which suggests that their compilers did not have photographic memories. Some of the lists include mnemonic devices, such as acronyms, to aid the user in recalling a list of words. Dotan (1971: col. 1424) reports an acronym designed to assist Bible students in remembering ‘the order in which the seven nations are listed’. An English correlate is the sentence people memorize to learn the order of the planets: each word in the sentence begins with the first letter of a planet’s name.

Masoretic lists indicating the phrases in which a word occurs in the Bible are introduced by the word ve-simanhon, which means literally ‘and their sign [is]’. Dotan (1971: col. 1424) translates the Hebrew word simanim as ‘mnemonic devices’. This implies that the Masoretes expected readers to recall all the occurrences of a word in the Bible, and that the phrases were provided only as an aid to memory. The fact that the word for ‘sign’ in ve-simanhon is in the singular suggests that the phrases are to be viewed as one long string

Figure 5. Page from a Hebrew Bible written in the Franco-German region in the early 13th century, prior to the compilation of the first Latin Biblical concordance. The Masoretic notes on the bottom record the occurrences of three words and phrases. The words are aligned on the right and are followed by a Hebrew letter indicating their frequency. The third column contains the abbreviation for the term meaning ‘and their mnemonic device is’. The last two columns record the phrases in which the words are found. Folio 98 recto of Ms Hunt. 11. By permission of the Bodleian Library, University of Oxford.
to be memorized, like the sentence about the planets. Early Masoretic lists have no punctuation between the phrases; later ones put a dot or extra space between them (see Fig. 5).

Access to Masoretic Bibles by Christian concorders

The annals of science document instances of simultaneous and independent discovery; the first Latin concordance could, in theory, have been ‘invented’ without knowledge of Hebrew Masoretic lists.10 There is, however, substantial evidence that the Latin concorders had access to Masoretic Bibles and actively sought them out.11

The first fact to be established is: were there Masoretic Bibles in France at the time that Latin concordances were being developed? (The work of the Masoretes began in the Middle East.) A search of the codicological12 database of the Hebrew Palaeography Project, called Sfar-Data (Beit-Arié, 1993: 106), yields quite a few dated Masoretic Bibles written in France before the compilation of the first Latin concordance. This fact in itself does not prove that the former influenced the latter; Hebrew Bibles could have languished in medieval libraries unconsulted by Christians (it has been found that 80 percent of books in a research library are never used). There is, however, extensive documentation of the study of Hebrew by Christians in France for a variety of purposes; negative ones include conversion of the Jews and censorship of Hebrew books.13 Study of the Bible in its original language in order to establish an accurate Latin text of the Old Testament, as well as commentaries on it, was a positive reason. As Beit-Arié (1993: 16) states: ‘That Christian scholars studied Hebrew, . . . drew from Jewish sources and even used Hebrew manuscripts, mostly biblical, is widely attested by many Christian exegetical texts’.

Exegesis of the Bible entails understanding word meanings. There was a great deal of interest in Hebrew language and grammar in France during the 12th and 13th centuries. Early Hebrew dictionaries were copied, and new works created. A codex in the Bodleian Library14 contains several works of this type; the supplement to the catalog15 identifies this manuscript as having been written in France in the mid-13th century, just when Latin concordances were being compiled. As was pointed out in Weinberg (1999: 114), early Hebrew dictionaries, dating from the 10th century, included Biblical passages to illustrate usage. Thus they were concordance-like structures.

That Hebrew Bibles were studied by Hugh of St. Cher (Loewe, 1969: 150), the director of the major early Latin concordance project, provides almost irrefutable proof of the influence of Masoretic lists on the development of the Latin concordance. This influence, however, is not recorded in publications describing the latter, notably the works of Richard and Mary Rouse (1979, 1990).

Directions for further research

The work of the Masoretes may have been influenced by contemporaneous linguistic activity of the Muslims, who were attempting to standardize the text of their sacred scripture, the Koran, in the 7th century (Pearson, 1986: 404–9). Scholars whom I have consulted consider this a plausible theory, considering that Hebrew grammarians learned from their Arab counterparts. Early Hebrew grammars and dictionaries were written in Judeo-Arabic (Brisman, 2000: ch. 1) but the 12th-century lexicographer Salomon Parhon is credited with transferring the information to Christian scholars by writing in Hebrew (Blau, 1971). Parhon’s dictionary was copied in France, and it is likely that the Latin concorders consulted it.

This article grants the possibility that the Masoretes learned from other cultures. It claims only that Hebrew Biblical word lists with phrase locators preceded Latin concordances by several centuries, and probably stimulated the idea for such indexes.

Another topic worthy of exploration is memory techniques in ancient cultures. Colleagues have suggested that Indian and Chinese techniques for memorization of large quantities of text be investigated. In a prior paper (Weinberg, 1999: 116) I quoted sources indicating that indexes begin when memory declines. Masoretic lists represent incredible feats of memory and, at the same time, extraordinary index structures.

Acknowledgments

My research in Israel and England was supported by grants from the Eugene Garfield Foundation and the Jesselson Foundation. St John’s University provided many types of support. This paper benefited from discussions and communications with experts on the Bible and codicology in the United States and Israel. The American scholars include Prof. David Berger (who kindly read a draft of the paper), Rabbi Miles Cohen, Dr Consuelo Dutschke, Prof. Sid Leiman (who kindly read the proofs) and David B. Levy. The Israeli scholars are Drs Robert Brody, David Lyons, Yosef Ofer and Jordan Penkower, as well as Prof. Malachi Beit-Arié and Tamar Leiter of the Hebrew Palaeography Project. The staff of the Institute of Microfilmed Hebrew Manuscripts of the Jewish National and University Library (Jerusalem), the Rare Book Room of the Jewish Theological Seminary of America (New York), the British Library (London), Cambridge University Library (England), and of the Bodleian Library (Oxford) are also gratefully acknowledged.

Notes

1. There is disagreement on the vocalization, and hence the transliteration, of the Hebrew word mesorah. Many English-language works use the spelling Masorah. This paper follows the preferred vocalization of Even-Shoshan’s dictionary (1997, vol. 3: 80) but retains the alternative spelling in quotations.

2. Bibliothèque Nationale Ms. no. Heb. 148; University of Halle ms no. Yb 4:10 I examined the microfilms of both versions in Jerusalem.

3. Dr Yosef Ofer provided several examples of references in Masoretic manuscripts to ordinal numbers of sedarim, e.g. ‘the 3rd seder of Chronicles’ (personal communication, May 2000).

4. The sequence of books in the Old Testament differs in Jewish and Christian traditions, and there is variation within these traditions. Dotan (1971: col. 1406) notes the differences found in Hebrew manuscripts.

5. The word et has no English equivalent; it comes before a direct object in Hebrew.
References


Weinberg, Bella Hass is a Professor in the Division of Library and Information Science, St John’s University, Jamaica, New York 11439, USA; fax: +1-718-990-2071. She is a past president of the American Society of Indexers and the 1998 recipient of its Hines Award. Among her specialisms are Hebrew grammar, transliteration and cataloguing.