REVIEW ARTICLE: MA'AGARIM – HEBREW DIACHRONIC CORPUS AND CORPUS QUERY SYSTEM*

Tsvi Sadan
Bar-Ilan University

Homepage of Ma'agarim

1. INTRODUCTION

Ma'agarim,[1] a project of the Academy of the Hebrew Language, is both a (national) diachronic corpus of Hebrew, primarily for its *Historical Dictionary of the Hebrew Language*,[2] as well as a query system for this specific corpus. It became freely accessible online in 2014 with support from the Israeli Prime Minister's Office and is available now in two language interfaces – fully in Hebrew and partially in English.

The present article will examine these two aspects of Ma'agarim, including

* I would like to express my sincere gratitude to Ayelet Harel (The Historical Dictionary Project, The Academy of the Hebrew Language, Jerusalem) and Doron Rubinstein (The Historical Dictionary Project, The Academy of the Hebrew Language, Jerusalem) for patiently sharing with me invaluable undocumented information about Ma'agarim. I would also like to thank Elena Luchina (Inalco, Paris) for her constructive comments and suggestions on an earlier draft of this article. Any errors and inaccuracies that remain, however, are my sole responsibility.

[1] http://maagarim.hebrew-academy.org.il/.

[2] http://hebrew-academy.org.il/topic/%D7%9E%D7%A4%D7%A2%D7%9C%D7%94%D7%9E%D7%99%D7%9C%D7%95%D7%9F/%D7%AA%D7%95%D7%9C%D7%93%D7%95%D7%AA-%D7%9E%D7%A4%D7%A2%D7%9C%D7%94%D7%9E%D7%99%D7%9C%D7%95%D7%9F-%D7%94%D7%94%D7%99%D7%A1%D%98%D7%95%D7%A8%D7%99/

its primary sources and linguistic annotation, and its concordance and automated assistance for dictionary writing (or lack thereof). This analysis will be undertaken both independently in comparison with major corpora in other languages as well as with Sketch Engine.[3] Sketch Engine is a powerful and innovative online corpus query system used by a number of world-renowned dictionary publishers, such as Oxford University Press and Cambridge University Press, and national dictionary projects, such as Trojina, Institute for Applied Slovene Studies.[4] As this article is not a user manual, it will not be able to touch upon all the technical details of Ma'agarim as both a diachronic corpus and a corpus query system. Special attention will be paid at the end of the article to the possible, unintended, use of Ma'agarim for language teaching and learning in addition to its intended use as the data source for the *Historical Dictionary of the Hebrew Language*.

The major corpora, including 1) national diachronic, 2) non-national diachronic, and 3) national synchronic ones, examined and compared with Ma'agarim in this article are: 1.1) Georgian National Corpus (Project);[5] 2.1) Corpus of Historical American English,[6] 2.2) Corpus of Historical Portuguese,[7] 2.3) Diachronic Corpus of Written Italian,[8] 2.4) Helsinki Corpus of English Texts,[9] 2.5) Historical Corpus of the Welsh Language,[10] and 2.6) Penn Corpora of Historical English;[11] 3.1) Bulgarian National Corpus,[12] 3.2) Croatian National Corpus,[13] 3.3) Czech National Corpus,[14] 3.4) Hellenic National Corpus,[15] 3.5) Hungarian National Corpus,[16] 3.6) National Corpus of Polish,[17] 3.7) Russian National Corpus,[18] 3.8) Slovak National Corpus,[19] and 3.9) Turkish National Corpus.[20]

Ma'agarim is unique in the world in that it is probably one of the few *national* and *diachronic* corpora that are already available online to the public and it covers such a long period of time – a feat unparalleled by most

[3] http://www.sketchengine.co.uk/.
[4] http://www.trojina.si/.
[5] http://gnc.gov.ge/gnc/static/portal/gnc.html.
[6] http://corpus.byu.edu/coha/.
[7] http://corporavm.uni-koeln.de/colonia/.
[8] http://corpora.dslo.unibo.it/DiaCORIS/.
[9] http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/.
[10] http://people.ds.cam.ac.uk/dwew2/hcwl/menu.htm.
[11] http://www.ling.upenn.edu/hist-corpora/.
[12] http://dcl.bas.bg/bulnc/.
[13] http://www.hnk.ffzg.hr/.
[14] http://www.korpus.cz/.
[15] http://hnc.ilsp.gr/.
[16] http://corpus.nytud.hu/mnsz/.
[17] http://www.nkjp.pl/.
[18] http://www.ruscorpora.ru/.
[19] http://korpus.juls.savba.sk/.
[20] http://www.tnc.org.tr/.

other languages and their respective diachronic corpora. It is also special in that such a corpus was already envisioned well before the start of the so-called "corpus revolution" in lexicography[21] in the 1980s. The whole corpus project became fully computerized only in 1995, the first fruit of the work was made public as a CD-ROM in 1998, and it made its first online appearance in 2005, then as a subscription-based service. The idea of compiling the *Historical Dictionary of the Hebrew Language*, however, was proposed much earlier, in 1937, before the establishment of the State of Israel. It was officially launched as a project in 1954, one year after the Academy of the Hebrew Language was legally recognized as the supreme body for research in the Hebrew language. The task of writing individual dictionary entries, however, has only recently begun, and in a very limited scale so far.

2. MA'AGARIM AS A DIACHRONIC CORPUS[22]

2.1 Primary Sources[23]

Ma'agarim at its present stage of development covers four of the five main periods in the long history of Hebrew language – that is, Hebrew of the Second Temple Period, Rabbinic Hebrew, Medieval Hebrew, and Modern Hebrew (until the establishment of the State of Israel), spanning about two millennia from ca. 200 BCE to 1948. The Hebrew Bible has not been included yet as it is already freely available electronically, such as with TanakhXL [24] or commercially as part of software packages such as Accordance [25] and BibleWorks. [26] In addition, a number of scholarly dictionaries already exist. Yet, work will start soon on this first important period of the language. There are several comprehensive monolingual dictionaries for Modern Hebrew, such as *Rav-milim,*[27] though none of them is corpus-driven or even corpus-based. The extent of Hebrew primary sources to be included in the corpus, especially from before the modern period, is not yet clear.

The primary sources included in Ma'agarim as of this writing (January 2016) can be divided as follows by period and genre in terms of their respective number and size in words (tokens):

---

[21] See, for example, Hanks (2012) for details.

[22] A historical corpus is not the same as a diachronic corpus. It can be either diachronic or synchronic. If it is only concerned with one historical period of the language in question, then it is a synchronic corpus; only if it covers multiple historical stages of the language is it diachronic. See, for example, Claridge (2008: 242) on this distinction as well as possible restrictions and problems. See also Brinton (2015) on the latter issues.

[23] See the following clip for guidance: https://www.youtube.com/watch?v=YSTTaRdXZms.

[24] http://www.tanakhml.org/.

[25] http://www.accordancebible.com/.

[26] http://www.bibleworks.com/.

[27] http://www.ravmilim.co.il/.

Period	Number of Sources	Size in Words (Tokens)
-200 – 0	304	60,786
0 – 300	227	825,307
300 – 600	583	1,883,062
600 – 800	354	864,985
800 – 1100	4,960	135,332
1100 – 1300	386	2,307,006
1300 – 1500	154	15,342
1500 – 1750	0	0
1750 – 1918	266	3,047,897
1918 – 1948	54	242,757
Total	7,289	ca. 9.4 million

Genre	Number of Sources	Size in Words (Tokens)
Belles Lettres	122	1,152,877
Bills	6	207
Dead Sea Scrolls	303	1,019,240
Epigraphy	241	4,716
Geonic literature	585	16,820
Karaite literature	63	678,885
Letters	531	24,304
Linguistics	16	386,403
Non-fiction and journalism	166	1,017,257
Piyyut and prayer	2,220	49,559
Rabbinic literature	30	10,020
Science	36	1,141,143
Spanish poetry	2593	230,001
Talmud and Midrash	377	3,651,042
Total	7,289	ca. 9.4 million

It may be too late to change this classification into genres; unfortunately, historical periods and geographic locations are mixed as part of genres. For example, "Geonic literature" and "Spanish poetry" are genres delimited by a historical period and a geographic location respectively.

It must be emphasized that since new primary sources are being constantly added to the corpus, these numbers only reflect what is available to the public at present. All the primary sources already included or to be included are typed in manually from their respective oldest manuscripts or first printed editions. In addition, since the work on some periods was started earlier than that on others, the corpus in its present stage of development does not necessarily reflect the putatively planned distributional balance of the primary sources from various historical periods; for example, the period between 1500 and 1750 is not represented yet in the corpus.

Clicking on the left tab of Ma'agarim's homepage retrieves a list of all the primary sources that match the search criteria. Such a list can be retrieved by searching a specific author or primary source, and the search can also be filtered by period and genre in the window to the right. One can even read the *whole* text of any primary source.

A comparison with major corpora in other languages, including national diachronic, non-national diachronic, and national synchronic ones, could clarify the quantitative characteristics of Ma'agarim. The following list shows the size of each of these corpora as well as its time span if it is a diachronic corpus:

1) National diachronic corpus
   1.1) Georgian National Corpus (Project):[28] 5th century CE - present; (in preparation)
2) Non-national diachronic corpora
   2.1) Corpus of Historical American English: 1810–2009; 400 million
   2.2) Corpus of Historical Portuguese: 16th century – 20th century; 5.1 million
   2.3) Diachronic Corpus of Written Italian: ?1750/1861 – 1945; (size unknown)
   2.4) Historical Corpus of English Texts: 730 – 1710; ca. 1.6 million
   2.5) Historical Corpus of the Welsh Language: 1500 – 1850; 420,000
   2.6) Penn Corpora of Historical English: mid-12th century – early 20th century; ca. 404,000
3) National synchronic corpora
   3.1) Bulgarian National Corpus: over 1.2 billion
   3.2) Croatian National Corpus: 216.8 million

[28] See, for example, Gippert & Tandashvili (2015) on the design principles of this new corpus, currently in the planning stage, which is one of the few national diachronic corpora together with Ma'agarim.

3.3) Czech National Corpus: over 2 billion
3.4) Hellenic National Corpus: over 47 million
3.5) Hungarian National Corpus: 187.6 million
3.6) National Corpus of Polish: 1.8 billion
3.7) Russian National Corpus: over 600 million
3.8) Slovak National Corpus: over 829 million
3.9) Turkish National Corpus: 50 million

Although it is not clear how much of the planned corpus has been completed, one thing seems obvious: even when Ma'agrim has been completed, it will be much smaller than many of these national synchronic corpora.

2.2 Linguistic Annotation[29]

The three major types of linguistic annotation of corpora are morphological annotation (also known as lemmatization), morphosyntactic annotation (also known as part-of-speech tagging), and syntactic annotation (also known as parsing). The choice of linguistic annotation also affects the types of corpus query possible and the accuracy of their results. For a morphologically rich and complicated language like Hebrew, the first type of linguistic annotation – that is, morphological annotation – is probably the most important.

The following list shows which of these linguistic annotations is used, as known, in the other corpora mentioned above (corpora with no publicly available information about their respective annotations were omitted):

2) Non-national diachronic corpora
2.2) Corpus of Historical Portuguese: morphosyntactic annotation
2.6) Penn Corpora of Historical English: morphosyntactic annotation, syntactic annotation
3) National synchronic corpora
3.1) Bulgarian National Corpus: morphological annotation, morphosyntactic annotation
3.2) Croatian National Corpus: morphological annotation, morphosyntactic annotation
3.3) Czech National Corpus: morphological annotation, syntactic annotation
3.5) Hungarian National Corpus: morphosyntactic annotation
3.7) Russian National Corpus: morphological annotation, morphosyntactic annotation, semantic annotation
3.8) Slovak National Corpus: morphological annotation

In least this limited list of existing corpora, morphological annotation is the most commonly used type of linguistic annotation, and, for some it is even

[29] See, for example, Gries & Bereez (forthcoming) for an excellent, concise survey of linguistic annotation in and for corpus linguistics, including corpus lexicography.

the only one. Ma'agarim systematically offers morphological annotation. It is not syntactically annotated at all, and, strangely, morphosyntactic annotation is optional – that is, it is up to each annotator to decide whether to annotate his or her data morphosyntactically. The corpus is annotated manually, and then it is carefully checked by three researchers in order to ensure accuracy. No publicly available tag set is used, though there is a plan to switch to TEI,[30] an international standard used by a number of corpus projects (as well as for other purposes).

3. MA'AGARIM AS A CORPUS QUERY SYSTEM

3.1 Concordance[31]

Kilgarriff (2013: 78) mentions the following aspects of a dictionary's creation which a corpus supports:

- headword list development
- for writing individual entries:
  - discovering the word senses and other lexical units (fixed phrases, compounds, etc.)
  - identifying the salient features of each of these lexical units
  - their syntactic behavior
  - the collocations they participate in any preferences they have for particular text-types or domains
  - providing examples
  - providing translations

Although, as a corpus, Ma'agarim can safely be assumed to be the main data source for the *Historical Dictionary of the Hebrew Language*, as a corpus query system, it does not support the first option of generating a list of all the word-forms and/or lemmata, unlike, for example, Sketch Engine.[32] It thus seems that dictionary writers have such a list in advance. The lexicographic approach is thus corpus-based rather than corpus-driven[33] in this area, and Ma'agarim, as a corpus query system, cannot be used either for some of the above mentioned aspects for writing individual entries, such as their syntactic behavior and collocations.

Ma'agarim offers six query types: 1) word, 2) free text, 3) root, 4) smart phrase, 5) patterns, and 6) adjacency. In all the query types, search texts must

[30] http://www.tei-c.org/.

[31] See the following clips for guidance:
https://www.youtube.com/watch?v=e8C-c8csvSk.
https://www.youtube.com/watch?v=NbAxr0Ax6tM.

[32] See Kilgarriff et al. (2014) on the evolution and present functionality of this sophisticated online corpus query system. See also Kilgariff et al. (2012) on the use of Sketch Engine as infrastructure for historical corpora, including their subtypes, diachronic corpora.

[33] See, for example, Hanks (2012) on this important distinction in lexicography.

be typed with no vowel signs. The first is probably the most basic query type, and it has three subtypes: a) a word in the sense of lemma, b) a word in the sense of word-form, and c) both. Querying the corpus in this type returns a list of possible lemmata disambiguated by vowel signs and senses. The third query type is straightforward.

Querying the corpus using the word ערב, for example, 28 lemmata and ten roots are retrieved in query types 1a and 3 respectively. Clicking the retrieved roots will lead to the lemmata. Clicking any of the lemmata displays a concordance. It follows, therefore, that reaching the desired lemma after typing a search term in query types 1 and 3 is a two or three step process respectively.

In the other four query types, a search term is composed of more than one word. Type 2) searches a verbatim combination of words and one wild card (* for any one character). Type 4) searches a more sophisticated word combination by specifying their distance and order, etc. Type 5 searches word-forms or lemmata according to their verbal or non-verbal templates or both. Type 6 is a collocation generator retrieving a list of words *immediately* preceding or following a search term. This query type could be of immense practical value, but its limitation to immediate adjacency also limits its usefulness.

Search and search results can be filtered by period, genre, author, and primary source. But more sophisticated queries with regular expressions or specifying parts of speech as well as advanced operators that are part of query options, which exist, for example, in Sketch Engine, are absent from Ma'agarim.

The display format of the concordance could also be more user-friendly. The present format may be convenient enough when one is to examine each occurrence of a lemma closely, but it can cause a serious problem when browsing for good dictionary examples from a long list of occurrences for a very common lemma. These occurrences should have been displayed in the standard KWIC (keyword-in-context) format with metadata hidden from the main window but viewable by clicking if needed.

3.2 Automated Assistance[34]

Since the actual dictionary writing using Ma'agarim as a corpus is in its infancy and is done only by one person, it may not be sufficiently apparent that the present user interface for its concordance forces potential dictionary writers to spend a lot of unnecessary time scrolling manually through lists of occurrences for lemmata. For example, the lemma בַּיִת has more than 40,000 occurrences spanning more than 1,500 pages. Scrolling through all these

[34] See, for example, Rundell & Kilgarriff (2011) and Rundell (2012) for an extensive list of automated functions assisting dictionary writing.

pages and checking all these occurrences manually is time-consuming and can very easily make one lose sight of the forest for the trees.

More and more operations previously handled by people in practical lexicography are being automated. This is also the case with corpus query systems. The most famous automated assistance to handling of a massive number of such occurrences is the so-called word sketch,[35] which is part of Sketch Engine. In the words of its developers, a word sketch is "a one-page summary of a word's grammatical and collocational behaviour"[36] and "can be seen as a draft dictionary entry."[37] In other words, "[t]he system has worked its way through the corpus to find all the recurring patterns for the word and has organized them, ready for the lexicographer to edit, elucidate, and publish. This is how word sketches have been used since they were first produced."[38] A similar function is of paramount importance for Ma'agarim, as a corpus query system, to serve prospective users-cum-dictionary writers efficiently.

Another welcome function to Ma'agarim as a corpus query system would be GDEX (Good Dictionary Examples), also part of Sketch Engine. Again, in the words of its developers, GDEX "works by sorting a concordance, so the corpus lines judged best by the algorithm are shown first. Then the lexicographer should not have to read many of them before finding a good one. The same core technique has also been used to score documents and to exclude low-scoring documents from a corpus entirely."[39]

However, even if Ma'agarim would be equipped with such automated assistance functions for dictionary writing, "[f]or the time being, the hardest parts of lexicography – word sense disambiguation, definition-writing (for monolinguals), and providing translation equivalents (for bilinguals) – still require expert intervention by skilled lexicographers. But none of these operations is inherently intractable."[40]

4. Possible Use of Ma'agarim for Language Teaching and Learning

Ma'agarim was conceived and has been being built as the data source for the *Historical Dictionary of the Hebrew Language*. In spite of the limitations in its corpus query system at the present stage of development, however, it can also serve teachers and learners of Hebrew, especially of the Second Temple period, the period of the Mishna and the Talmud, the Middle Ages, and the

[35] Word sketch can be experienced in the free pedagogical lite version of Sketch English called Sketch Engine for Language Learning (SkELL): http://skell.sketchengine.co.uk/.
[36] Kilgarriff et al. (2014: 9).
[37] Kilgarriff et al. (2014: 10).
[38] Kilgarriff et al. (2014: 10).
[39] Kilgarriff et al. (2014: 29).
[40] Rundell (2012: 28–29).

modern era until the establishment of the State of Israel.

Thomas (2015) demonstrates with ample examples how to tap the full potential of Sketch Engine, which is primarily a corpus query system but is preloaded with many huge corpora in multiple languages, including a few diachronic corpora. Similar things may be possible with Ma'agarim for discovering the above-mentioned periods of Hebrew, though with limitations.

Ma'agarim can be primarily used as an online repository of primary sources in Hebrew meticulously selected by competent linguists. Unlike in many other online corpora, reliable full texts – and not only lines containing search terms – are freely accessible. This possibility alone makes it worth introducing Ma'agarim to the classroom. The currently available query types and functions are more than adequate for this purpose, if not for the purpose of dictionary writing. Ma'agarim can also be used as a concordance, especially for those periods of Hebrew for which there is no dedicated modern scientific dictionary. It is, however, still limited in functionality to be used as a collocations dictionary. The use of Ma'agarim for language teaching and learning is restricted to morphological and lexical levels. As it is not annotated syntactically, it cannot be used for checking, for example, syntactic patterns and other syntactic behaviors of the queried lemmata.

5. SUMMARY

All in all, Ma'agarim is a significant milestone in the history of Hebrew corpus linguistics and prospective corpus(-based) lexicography. It consists of two parts: 1) a (national) diachronic corpus of Hebrew from the Second Temple period to the establishment of the Stata of Israel with the best primary sources carefully chosen and checked by experts at the Academy of the Hebrew Language, and 2) its corpus query system.

As a corpus, Ma'agarim is already quite impressive. However, there is still one important period of language – Biblical Hebrew – that is missing, not enough primary sources have been included in certain other language periods, and the part on Modern Hebrew does not go beyond the year of Israel's independence, thus omitting more than half of the history of Modern Hebrew.

As a corpus query system, Ma'agarim still leaves much room for further improvements. Since the actual process of dictionary writing has just begun, at a very limited scale, its limitations may not be so keenly felt now. Once this task is underway more seriously, both quantitatively and qualitatively, the need will become more compelling. The corpus query system can hopefully be improved to meet the demands of dictionary writers, facilitating their daunting task as, for example, Sketch Engine already does for a number of world-renowned dictionary publishers and national dictionary projects.

Ma'agrim seems to be unique as a functioning national diachronic corpus

and corpus query system. Other such national or international projects to build diachronic corpora, with or without compiling historical dictionaries, include the Doha Historical Dictionary of Arabic[41] and Georgian National Corpus (Project) respectively. Both of them are, however, still in their planning stage, while Ma'agarim was already conceived decades ago.

REFERENCES

Claridge, C. 2008. Historical Corpora. In: A. Lüdeling & M. Kytö. (eds.). *Corpus Linguistics: An International Handbook* 1. Berlin: Walter de Gruyter. 242–259.

Gippert, J. & Tandashvili, M. 2015. Structuring a Diachronic Corpus: The Georgian National Corpus Project. In: J. Gippert & R. Gehrke (eds.). *Historical Corpora: Challenges and Perspectives*. Tübingen: Narr. 305–322.

Gries, S. T. & Berez, A. L. Forthcoming. Linguistic Annotation in/for Corpus Linguistics. In: N. Ide & J. Pustejovsky (eds). *Handbook of Linguistic Annotation*. Berlin: Springer.

Hanks, P. 2012. The Corpus Revolution in Lexicography. *International Journal of Lexicography* 25: 398–436.

Kilgarriff, A. 2013. Using Corpora as Data Sources for Dictionaries. In: H. Jackson (ed). *The Bloomsbury Companion to Lexicography*. London: Bloomsbury. 77–96.

Kilgarriff, A. et al. 2012. The Sketch Engine as Infrastructure for Historical Corpora. In: J. Jancsary (ed.). *Empirical Methods in Natural Language Processing: Proceedings of the Conference on Natural Language Processing 2012*. Vienna: ÖGAI. 351–356

Kilgarriff, A. et al. 2014. The Sketch Engine: Ten Years On. *Lexicography* 1: 7–36.

Rundell, M. 2012. The Road to Automated Lexicography: An Editor's Viewpoint. In: S. Granger & M. Paquot (eds.). *Electronic Lexicography*. Oxford: Oxford University Press. 15–30.

Rundell, M. & Kilgarriff, A. 2011. Automating the Creation of Dictionaries: Where Will It All End? In: F. Meunier et al. (eds.). *A Taste for Corpora: In Honour of Sylviane Granger*. Amsterdam: John Benjamins. 257–281.

Thomas, J. 2015. *Discovering English with Sketch Engine*. N.p.: Versatile.

[41] http://www.dohadictionary.org/.