Rav-Milim – a Modern Dictionary for an Ancient but Thriving Language

Yaacov Choueka

1. Hebrew is probably one of the oldest languages in current usage, and its dictionary-making history goes back more than a thousand years. It is not our intention here to trace the whole history of Hebrew dictionaries or lexicographic compendia, but rather – as a contrastive background to this brief presentation of the *Rav-Milim* dictionary of Modern Hebrew (hereafter MH) – to mention the modern ones, i.e. those in vogue in the twentieth century before the *Rav-Milim* publication in 1997.

By universal opinion, the scene for the Hebrew dictionaries in the twentieth century was dominated by three major and influential works. First and foremost is the Eliezer Ben-Yehuda 26-volume dictionary, a monumental work of erudition and scholarship by "the reviver of the Hebrew language", the publication of which started in 1908 but ended only in 1959. This is an OED-type of historical dictionary, whose glossary included not only (though mostly) Biblical and Rabbinical terms, but also whatever MH ones were available then, especially those coined by Ben-Yehuda himself. The Gur dictionary (1934-36) was really the first general dictionary of MH, quite popular in the late thirties and forties. Finally, the enormously popular Even-Shoshan one – first published in 1947-52, then reprinted countless times in various formats and numbers of volumes (with only one major revision in 1970) – completely dominated the scene in Israel for almost 50 years, being present, virtually, in most local households.

To complete this picture, one should mention three other dictionaries whose impact was rather negligible: Cnaani's 18-volume dictionary (1960-82), Alcalay (1969-71) and Medan (1954).

All in all, then, just six dictionaries in a whole century, one of them – updated and revised in only minor ways – exclusively dominating the scene for most of the second half of that century, and none of them with any computerized components. Thus, during a period when not only the State of Israel and the Hebrew language were undergoing extraordinary dynamic cycles of changes and expansion, but the whole world was – and still is – exploring new frontiers (and devising new terms and semantic fields to describe them) in technology and science, and in intellectual and social life, dictionary making in MH was in practice frozen for some fifty years.

This was the state of affairs in late 1992, when it was decided to compile and publish – both in print and in electronic form – a new and up-to-date illustrated dictionary of MH, *Rav-Milim* [*Master-Words*], with a shorter companion – richly annotated and copiously illustrated in color, specially adapted to young children and teenagers in elementary and secondary schools – *Junior Rav-Milim*. Here we restrict ourselves to the description of the unabridged *Rav-Milim* printed version, its underlying philosophy and some of its salient features.

2. Although not a purely corpus-based dictionary, the *Rav-Milim* design was deeply influenced by computerized methodologies and techniques of natural language processing developed since the mid-1980s, not only in its production and in its extensive cross-checking algorithms, but also in its very structure and editing method. Indeed, since the late eighties, computers have altered the way we view dictionaries, their functionality, their aims, and the degree of thoroughness, coverage, accuracy, precision and methodical writing we have come to expect from them. These influences were masterly described by Krishnamurthy in a previous issue of this newsletter (2002). True enough, the Krishnamurthy paper was about EFL dictionaries; taking into account, however, that for a great majority of the population of Israel, immigrants from all over the world, Hebrew is indeed, to a certain extent at least, a "foreign language", such insights are highly relevant to a general dictionary of MH as well.

From its very inception, it was decided that *Rav-Milim* (RM) will be developed along completely different – in fact, radically different – lines than previously published dictionaries of Hebrew (PPDH in short), constituting an "anti-thesis" – so to speak – to them on almost each and every methodological issue of dictionary designing and editing. It differs from PPDH in the list of entries, in the entry's structure, in the entry's "explanation", in the detailed and fine analysis of the various meanings of the entry and their order, in the usage examples, in the usage directives,

Yaacov Choueka was born in Cairo in 1936, immigrated to Israel in 1957, and got a PhD in Mathematics from the Hebrew University in Jerusalem in 1971. His computational work on the Hebrew language has pioneered innovations on a global scale, including the development of a computerized lemmatizer and morphological analyzer (1964), leading the Responsa corpus project (1974-86) for a full-text system for Rabbinical literature, and the Rav-Milim project (1989-97) for laying the infrastructure of advanced processing of Hebrew, producing, among others, *Rav-Milim* dictionary and Nakdan computerized program for automatic vocalization of texts. He was twice recipient of the annual award of the Israel Association for Information Technology, and received the Prime Minister Prize for computing in 1997. Since 1964, Professor Choueka has been associated with the Department of Mathematics and Computer Science at Bar Ilan University, where, since 1974, he heads the Institute for Information Retrieval and Computational Linguistics.
ycsarah@netvision.net.il

Rav-Milim

a Comprehensive Dictionary for Modern Hebrew
Yaacov Choueka and the Rav-Milim team
C.E.T., Miskal and Steimatzky, Tel Aviv, 1997
6 vol. 16+1955 pages
ISBN 965-448-323-8

***Rav-Milim* Online**

The online version of *Rav-Milim* (www.ravmilim.co.il), developed and maintained by Melingo Ltd, is the only full Hebrew dictionary on the web. It offers several features that are not included in the printed version.

Morphological analysis

This is a unique feature of the online version, which is able to identify the correct lexical entry of any word, even if it has multiple inflections. For example, the user can look for the meaning of *vayashkuhu* [*and they will water him*], without knowing the root or basic form of the word, and find the correct lexical entry – *hishqa* [*watered*] – with its explanation, translation and a full grammatical analysis of the inflected form. For Hebrew this is a critical feature, because prepositional proclitics

in the registers' annotations, in the "etymological" notes, and in the thorough and detailed processing of collocations and when, where and how to include them. At the risk of being somewhat simplistic, we can state schematically that RM is intended to be synchronic and not diachronic, descriptive and not normative, explanatory and not definitional, contemporary and not archival, illustrative and not quotation-minded. Furthermore, a maximum of uniformity and consistency in the dictionary compilation was assured (and continuously checked by the computer) by having all editorial questions discussed, decided and recorded formally by the editorial committee, which counted among its members five prominent professors of Hebrew.

In the following we shall briefly present the main features of RM, most of which were "firsts" in Hebrew lexicography, and some of which have since been adapted in a few Hebrew dictionaries that were published after it.

3. The written form of Hebrew – as that of other Semitic languages – is an essentially unvocalized one, vocalization being marked by diacritical points that may appear below, above, or inside the word's letters. Such a vocalization is however rarely used in everyday writing, except for Biblical texts, poetry or (more recently) for children's books. To alleviate some of the annoying ambiguity that would thereby result in many different "readings" of a given word, it has been customary to add in appropriate positions of the word some mater lectionis: *Vav* for the vowels O and U, *Yod* for E and I, and *Aleph* for A, thus producing the so-called plene script. Still, most PPDH were edited in the formally vocalized grammatical script, and the entries were also given – and therefore sorted – in this form. We thought that such a vocalized script would seem totally out of context to any reader who never encounters such texts elsewhere, not to mention its childish (on one hand) and somewhat paternalistic (on the other hand) projection. RM is therefore edited in the plene spelling, and the headwords are given in that script, since this is exactly how the user will usually see it in a publication and look for it in the dictionary. Following the plene headword, its grammatical vocalized form is given, so as to assist the user in pronouncing it correctly and recognizing its pattern. Additionally, a pointer is given from that form, in its alphabetical position, to the plene one, just in case the user encounters that form or is extrapolating from the given plene one and looking for it in RM. Incidentally, the number of spelling variants in Hebrew is rather large, also because of different ways of transcribing loan words from many languages over the ages, whether from Aramaic in ancient times or mainly from English most recently; since having pointers from these variants in the main dictionary page would have hopelessly encumbered it (indeed many pages in PPHD consist mostly of such pointers!), all pointers pertinent to a given page were collected and printed in a separate section at the bottom of that page.

The list of entries in RM is distinguished both by what it contains and by what it omits. Besides listing virtually every (Hebrew) Biblical word and most terms from early Rabbinical sources (except *Hapax Legomena*, whose meaning is not well understood and is inferred only from the context), the list contains every word in current usage, from all registers – from the highest literary ones to the most colloquial and vulgar ones. The only criterion for inclusion was whether such an utterance can be read or heard somewhere; if so, then we must help the user understand it, by including it and its meanings in the dictionary (this was indeed the first time ever that such terms were included in a general dictionary). On the other hand, the word's register is always clearly marked; from the highly literary (to warn the reader against using such a word in – say – asking directions) to the colloquial, vulgar or obscene, as well as corrupt form of, etc.

We included as entries also utterances that are not, linguistically, "words" of the language, but are used in certain ways specific to Hebrew, such as *tsvits tsvits* for denoting a bird song, *miaou* for a cat call, *koukourikou* for the rooster call, *sha* for requesting silence, etc.

Special consideration had to be given to the inclusion of "encyclopedic" terms and knowledge, and of terms from various scientific and technological domains. A dictionary is neither an encyclopedia nor a complete guide to the fauna and flora of the world or even of a certain region of it. As a rule-of-the-thumb, any term that may potentially occur in a general publication was included, and any term that occurs only in the relevant professional publications was excluded.

For various types of "non-linguistic" terms, the decision on whether to include them as entries in RM was made by the editorial committee, and rigorously implemented. Following are some examples of such decisions:

- No proper name of anyone (living or dead) is to be included; literary or

mythological figures are mentioned to the extent that they are used metaphorically (Samson, Venus, Casanova) or in collocations (Richter's scale, Columbus egg).

- Country names are included, along with the language(s), capital and up to three cities, and two denominations of currency – the minimal one and the main one (cent and dollar, penny and pound).
- Places in Israel are included if they have more than 5000 inhabitants as per the last Israel census.
- No specific "creations" (books, theater, arts, etc) are included, with the exception of the 24 books of the Bible and the canonical early Rabbinical sources.
- All elements of the cyclical table are included, with a uniformly designed explanation.

On the other hand, we omitted from RM thousands of obsolete entries that appeared in PPHD: words coined from the late nineteenth century and loan words from other languages that were almost never used, even words officially coined by the Academy of the Hebrew Language that did not enjoy wide acceptance, etc. Our policy was that not every word used once or twice by a writer, as great as he or she may be, should be automatically recorded in the dictionary. Delicate editorial considerations sometimes have to be applied in such cases.

Another issue that well illustrates the spirit of RM is the following. Because of the peculiar history of the Hebrew language, many words have persisted and are in current usage in certain conjugated or derived forms, while the original variant is – and was – never in use (*hav* [*give*], only in the imperative; *be'etyo* [*because of him/it/that*], only with the preposition and the pronoun). PPHD used, in such cases, to "extrapolate" and invent the presumed original form and list it as a dictionary entry. We refrained from inventing words, and such terms were given as entries "as are", which is anyway the form in which the user will encounter such words and look for them in the dictionary.

4. "The principal reason for the existence of a general monolingual dictionary is its definitions. All the art and all the scholarship and all the scientific methods that the editors can command are required to study meanings and write definitions" (Gove, 1961).

Contrary to Gove's wise dictum, one cannot but notice that in most PPHD this aspect of dictionary compilation has been quite neglected, usually with the justification of offering one or more synonyms of the entry. In RM, however, we fully endorsed this statement, with all its consequences and ramifications, except, maybe, for replacing "definitions" by "explanations", since our aim was not to give an Aristotelian definition of an entry, but to explain it completely and precisely. According to the RM concept, the ultimate test of a good explanation is whether a user who has never encountered the word before can now understand it as fully and precisely as possible. On the one hand, we painstakingly analyzed and checked every word in the explanation to assure its appropriateness and pertinent coverage. On the other hand, we aimed at detailing explicitly all the nuances and shades of the basic meaning of the entry, as manifested in the different contexts in which it actually occurs. Indeed, as stated by Firth (1957), "you recognize a word by the company it keeps".

One example should suffice to clarify this approach. The adjective *ham* [*hot*, *warm*] is defined in Even-Shoshan only as "having a more or less high temperature". In RM, this entry details some 11 different meanings or usages in various contexts (that may well translate into different words in other languages), which an innocent reader would not be able to guess on her/his own. Thus, besides the basic meaning as in "hot soup" (vs. "cold soup"), we have "hot news" (but not "cold news"), "hot temper", "warm heart" (the former with a negative connotation, the later with a positive one), "warm voice" (specific voice texture), "warm clothes" (the clothes themselves are not warm, they warm the body), "he is hot" (which doesn't mean he has "a more or less high temperature", he is not sick, he just feels hot and would like to open the window), etc. Even the "Hot! Hot!" call in the hide-and-seek children's game deserves and gets its own numbered meaning. Indeed, the fine analysis of the extremely rich spectrum of the nuances of almost every word, according to the contexts in which it appears, is one of the greatest achievements and benefits of the application of computers to the processing of large corpora, and the lexicographer's efforts for collecting, classifying, sorting and adequately explaining these nuances is probably the most exciting and satisfying part of the dictionary making process.

When the meanings of an entry have changed throughout its history, they were always ordered in PPHD, traditionally, chronologically. In RM, which has always had the user in mind, meanings are ordered by decreasing frequency; the most frequent sense given first, and adequate period labels attached when necessary.

►

and pronominal enclitics are attached to the word, resulting in many different forms for the same word. In addition, the lexical entries appear with diacritical vowels (*niqud* [*pointing*]), so in case of an ambiguous word like *SFR*, it is easy to find out whether we are looking for *sefer* [*a book*], or *safar* [*counted*], or *sapar* [*a barber*], or *sfar* [*borderland*], etc. These features are particularly helpful for children, new immigrants and other learners, who do not know where to look for a word in a printed dictionary, especially when it is an inflected form.

Thesaurus

The thesaurus provides a rich selection of synonyms. Each synonym is linked to its own entry, making it very easy to 'click' one's way around the dictionary, travelling from a word to its synonyms, on to their definitions, and onwards through a wealth of linguistic information.

Bi-directional Hebrew/ English translation

The addition of the Hebrew-English and English-Hebrew translation provides users with the ability to translate any word. This, together with the fact that each Hebrew entry appears with vowels, makes the dictionary a useful translation resource.

Regular update

A great advantage of any online dictionary, and certainly of *Rav-Milim*, is the ability to constantly update it. Melingo continues to maintain and enhance

►

the website, and regularly adds new words that are released by the Academy of the Hebrew Language, such as *hetpes* [*stereotype*] and *dvekanut* [*perseveration*], or neologisms, including the latest slang new technological terms, such as *netiqa* [*netiquette*], or the latest slang, such as 'and *'en matsav* [*No way!*]. Not less important is the interactive nature of the online dictionary, i.e., it allows users to suggest new words, or to ask for clarifications about certain examples and illustrations, etc.

Phrases and idioms

The online version has two ways of finding a phrase in the dictionary: by looking at the list of phrases that appears under each component of the phrase, or by typing the whole phrase, a part of it or an inflected form of it. For example, the idiom *naga le-libo* [*touched one's heart*] can be found under *naga* [*to touch*] and *lev* [*heart*], as well as by typing any of the forms or inflections *nag'a le-libi* [(*she*) *touched my heart*], *nog'im le-libchem* [(*they*) *touch your hearts*], *yig'u le-libah* [(*they*) *will touch her heart*], etc.

Additional features

The website includes an **automated rhyming system** that presents all rhyming words for each lexical entry and sorts them according to the rhyme quality. A **crossword solving** feature enables the user to insert the known letters and number of letters in the word, and to receive a list of possible words from which the correct answer can be chosen.

Finally, an explanation is almost always followed in RM by one or more examples of usage, which only rarely are quotations from canonical writing. In nearly all cases, examples were carefully crafted to add interesting and useful details to the explanation.

5. One of the impacts of large corpora processing on linguistic studies in general, and on dictionary making in particular, since the mid-eighties, has been the recognition of the critical importance of collocations in defining the language elements and structure. If this is true for European languages, how much more so for Hebrew! Indeed, with the world dynamically revolving around us, the Hebrew language has constantly had to acquire and absorb numerous new words from the various domains of modern life activities. Although some new terms are adapted as loan-words "as is" and easily become part of current Hebrew, in many cases, however, Hebrew – being a Semitic language with a structure of 3- (or 4-) letter roots and derivation patterns – is quite resistant to such assimilation. A common productive solution is to have a two- (or three-) word Hebrew sequence to represent a new concept. A large number of single-word nouns in English, for example, such as *school*, *hospital*, *lawyer*, *accountant*, are represented in Hebrew by a two-word sequence.

In spite of that, the treatment of collocations in PPHD has been rather poor, to say the least. Very few collocations found their way into these dictionaries; phrasal collocations, idioms and even proverbs (!) were all mixed up; no clear guidelines were respected in terms of where and how to have the collocation's main entry (in fact, in an extreme example, a 4-word collocation actually appeared in 4 different entries with 4 different explanations!), or in terms of how to deal with, and uniformly represent, the "empty places" in some of these collocations, etc.

Having researched the problem of collocations already in the eighties (see 1983, 1988), I was strongly biased in favor of a comprehensive, systematic, rigorous and consistent treatment of the collocational part of RM. A small sample of the new features introduced in this endeavor now follows.

- To the question of when does a sequence of two or more words deserve its own entry in the dictionary as a collocation, a common answer is: when the meaning of the sequence is not the total sum of its components' meanings, and cannot be guessed from it. This is indeed an important criterion, but it is far from being unique. We delineated 12 different criteria that can justify such an inclusion, and every potential collocation was tested accordingly.
- Almost 10,000 new collocations were added in RM that never appeared before in PPHD. This is an extremely high figure when taking into account that the total number of (single-word) entries in PPHD is of the order of 35,000 entries only.
- Proverbs (e.g. 'not all that glitters is gold') were completely banned from the dictionary; phrasal collocations and idioms were sharply separated.
- Strict rules were set up and followed on where to introduce the main entry of a collocation and its explanation. The explanation appears, of course, only once, but pointers to that occurrence are given from every word of the collocation.
- Collocations were tagged by part-of-speech tags: nominal, verbal, adjectival, adverbial, etc. When necessary, morphological variants were added.
- Possible additions, omissions, replacements, etc, in the collocation text were marked clearly, in a uniform way.

With these steps and more, we indeed believe that the collocotional component of RM has made an important contribution to the clarification and systematic study of collocations in Hebrew.

To sum up: RM was a bold step taken to bring modern methodologies, trends and techniques to Hebrew dictionary making, applying overwhelmingly a computerized approach to its compilation and checking procedures. We believe that it has thus set a new standard of precision, coverage, methodology and systematization that will be hard to ignore.

Acknowledgements

Rav-Milim was the product of a large team of dedicated and competent professionals whom I have had the good fortune of leading and directing. I thank them all for the wonderful job they did, and I wish I had the space to mention them all by name. I would like still, at least, to thank personally Uzi Freidkin, Dr Haym Cohen, Yoni Ne'eman and Sara Choueka. May they all be blessed in their future endeavors.

Dictionaries and articles

Alcalay, Reuven. 1969-1971. *Milon 'Ivri Shalem* (*The Complete Hebrew Dictionary*). 3 vol. Ramat-Gan: Massada.

Ben-Yehuda, Eliezer. 1908-1959. *Milon ha-Lashon ha-'Ivrit ha-Yeshana ve-ha-Hadasha* (*A Complete Dictionary of Ancient and Modern Hebrew*). 16 vol.

Cnaani, Yaacov. 1960-1982. *Otsar ha-Lashon ha-'Ivrit* (*Treasure of the Hebrew Language*). 18 vol. Ramat-Gan: Massada.

Even-Shoshan, Avraham. 1947-1952. *Ha-Milon he-Hadash* (*The New Dictionary*). 5 vol; 1966-1970, 7 vol; 1997, 5 vol. Jerusalem: Qiryat Sefer.

Grazovsky (Gur), Yehuda. 1934-1936. *Milon ha-Safa ha-\Ivrit* (*Dictionary of the Hebrew Language*). 3 vol; 1947, 1 vol. Tel Aviv: Dvir.

Medan, Meir. 1954. *Me-'Aleph 'ad Tav – Milon 'Ivri Shimushi* (*From A to Z – a Practical Hebrew Dictionary*). Jerusalem: Achiasaf.

Choueka, Y., S.T. Klein and E. Neuwitz. 1983. 'Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus'. *ALLC Journal*, 4.34-38.

Choueka, Y. 1988. 'Looking for needles in a haystack or: locating interesting expressions in large textual databases'. *Proceedings of the RIAO Conference (Cambridge, MA)*. 609-623.

Firth, J.R. 1957. 'A Synopsis of Linguistic Theory, 1930-1955'. S*tudies in Linguistic Analysis*, 1-32. Oxford: Blackwell.

Gove, P.B. 1961. 'Linguistic Advances and Lexicography'. *Word Study*, October 1961, 3-8.

Krishnamurthy, R. 2002. 'The Corpus Revolution in EFL Dictionaries.' *Kernerman Dictionary News*, 10.23-27.

Milon Kis-Ariel

Maya Fruchtman

Milon-Kis Ariel [*Ariel Pocket Dictionary*] was compiled in response to concerns of teachers and families from the south of Israel, who wanted an up-to-date, learner-friendly Hebrew dictionary, which will be aimed explicitly at present-day young pupils, including new immigrants. This need arose because the existing school dictionaries were generally viewed as out-of-date, offering sloppy definitions, using archaic language, and lacking current words and meanings.

To achieve the goal of creating this Modern Hebrew learning tool, we established a competent editorial team consisting of experienced educators and linguists as well as professional specialists. In cooperation with a group of teachers, parents and students, we studied the specific requirements from such a dictionary, in relation to what was really necessary for the pupils at school and at home. The school dictionary was conceived as a first step in a larger project, having its own lexicographic database sources developed in relation to the target audience.

We decided to focus, in particular, on the following subject areas: flora and fauna, youth life, sports, civics and state institutes, communications, computers and technology, geography and history (particularly of Israel), literature, Judaism, loan words, current events and economics. We wanted to include old words besides modern ones, but to define them briefly and clearly, in a modern way.

The dictionary is a arranged in a straightforward alphabetical order, according to the first letter of the word, no matter what the Hebrew root is. The verbs are presented in the traditional way – based on the model of the past tense masculine singular. The entries include only definitions, not examples of usages, but they are often exemplified by the sub-entires. Vocalized spelling is used, in accordance with the rules of the Academy of the Hebrew Language. There are illustrations, for further emphasis or clarification of certain entries, and in order to make the book more lively and attractive.

The total number of entries includes 14,000 words and 4,000 phrases, of which as many as 3,000 appear for the first time in a Hebrew dictionary, for example: *Alzheimer's (disease)*, *anorexia*, *Intifada*, *Eurovision*, *Druzi* [*Druse*], *hashman* [*cardinal*], *divkit* [*sticker*], *hor ba-'ozon* [*hole in the ozone*], *duty-free, telemarketing, tampon*, *home'opatia* [*homeopathie*], *alpaka*, *hetsion* [*median*], *štrudel* [the symbol '@', officially *qrukhit*].

Maya Fruchtman studied Hebrew at Tel Aviv University and teaches at Bar Ilan University. Professor Fruchtman is the editor of the *Ariel* dictionary and lexicographic works, and has written extensively on the Hebrew language.
frucht@zahav.net.il

Milon-Kis Ariel

Maya Fruchtman
Kor'im, Qiryat Gat, 2001
1030 pages
ISBN 965-515-000-3