О КОРПУСЕ КАК ОБ ИНСТРУМЕНТЕ ИЗУЧЕНИЯ СОВРЕМЕННОГО ИВРИТА: ОГРАНИЧЕНИЯ И ПОТЕНЦИАЛ В ЦИФРОВУЮ ЭПОХУ

М. Е. Алексеева

Санкт-Петербургский государственный университет,
Университетская набережная, 11, 199034, г. Санкт-Петербург, Россия,
m.e.alekseeva@spbu.ru

Статья анализирует накопленный опыт использования корпусных инструментов и методик для изучения современного иврита. Автор классифицирует существующие корпуса, указывая, с одной стороны, на ограничения, а с другой, на потенциал их использования при решении различных лингвистических и прикладных задач. Рассматривается также опыт создания корпусов и результат применения квантитативных методов в синтаксических исследованиях.

Ключевые слова: корпус; корпусный анализ; современный иврит; синтаксис.

CORPUS AS A TOOL FOR STUDYING MODERN HEBREW: LIMITATIONS AND POTENTIAL IN THE DIGITAL AGE

M. E. Alekseeva

St Petersburg University,
Universitetskaya embankment, 11, 199034, Saint-Petersburg, Russia,
m.e.alekseeva@spbu.ru

The article analyzes the experience of using corpus tools and techniques for studying Modern Hebrew. The author classifies existing corpora, pointing out, on the one hand, the limitations, and on the other, the potential for their use in solving various linguistic questions as well as practical issues. The experience of creating corpora and the result of using quantitative methods in syntactic research are also considered.

Keywords: corpus; corpus analysis; Modern Hebrew; syntax.

The use of corpus tools and automated data analysis methods is gradually becoming more and more popular in modern linguistic research. The undoubted advantage of the corpus-based method is that such research is always “based on recorded speech samples, and not on made up examples, on segments of texts linked by certain semantic relationships, and not on isolated sentences” [1, p. 10]. Thus, corpus analysis makes it possible to identify relevant factors with greater efficiency and statistically assess the strength of their influence on the choice of a particular linguistic expression [2, p. 120–121].

Corpus analysis, for various reasons, is most actively used for studying major European and some oriental languages (in particular, English, Spanish, Chinese, Arabic and Japanese). However, in the last two decades, the number and volume of other languages corpora, including Modern Hebrew, has also rapidly increased. On the SketchEngine platform (sketchengine.eu) alone, users can access twelve Modern Hebrew corpora, some of which are available in multiple versions.

The goal of this article is to demonstrate the potential of corpus analysis for solving various theoretical and applied problems in linguistic research using the Modern Hebrew data, as well as to point out some limitations in using such technologies.

The issues involving the processing of natural languages that differ significantly from the Central European standard were demonstrated by many authors. For Hebrew, they were comprehensively described by Sh.Wintner in 2004 [3]. Among them, the phonological and morphological ambiguity of the Hebrew text, the problem of multiple interpretations of the syntactic structure and the variability of text division. Together with the relatively small number of speakers these factors slow down the development of Hebrew-based corpora and other resources significantly.

However, over the past twenty years, due to the efforts of researchers [4, 5, 6, etc.] and technological advances, the range of Hebrew-language corpora has been expanding, and the tools for quantitative analysis continue to improve constantly.

Currently, there exist both general (Hebrew General Corpus [7], 150 million words) and specialized corpora for various tasks: corpora of spoken Hebrew (Corpus of Spoken Israeli Hebrew [8] (18 hours), Map Task Corpus of the Open University of Israel [9] (32 units), Haifa Corpus of Spoken Hebrew [10]), Internet corpora (heTenTen21 [11] (2.7 billion words)), literary corpora (Gutenberg Corpora 2020 [12] (158 thousand words), Hebrew Drama Corpus [13] (950 thousand words)), corpus of children's speech CHILDES Hebrew Corpus [14] (807 thousand words), as well as parallel corpora (for example, Hebrew Translation Corpus [15] (1.1 million words)).

Among the listed there are both annotated and non-annotated corpora. Spoken language corpora, such as the Map Task Corpus [9], in particular require a complex tagging system [16], while for the multi-billion heTenTen21 Internet corpus [11] even the annotation does not completely disambiguate the contexts.

Despite some of the limitations that working with such corpora impose on researchers, corpus analysis is gradually becoming common practice among researchers of Modern Hebrew. Spoken language corpora are especially widely used [for example, 17, 18]. Using different corpora studies focus on lexical stress patterns [16], discourse markers [19, 20, 21] and aspectual tense system

of the verb in spontaneous speech [22], as well as stylistic features and discourse strategies in written text [23], and many other relevant and diverse topics.

Automated systems for extracting data from large text corpora are successfully used not only to study phonology, lexicon and morphology, but also the syntax of Modern Hebrew.

In particular, the author conducted a study of asymmetric object marking for Modern Hebrew, based on two research corpora constructed for this particular task: Hebrew Objects General Corpus (HOG corpus) and Hebrew Objects Targeted Corpus (HOT Corpus), with a total volume of about 101 thousand words. The dataset for these corpora was based on the annotated version of the heTenTen21 Internet corpus [11]. The fact that the aforementioned corpus provided general Part-of-Speech tagging, as well as was specifically annotated for Hebrew (NNT (Construct state noun), AT (Accusative marker ’et), POS (Possessive preposition šel and accusative marker ’et with a pronominal suffix), DEF (Definiteness marker a -), etc.), made it possible to automatically sort and categorize contexts that are most relevant to the objectives of the study, resorting to manual sorting only at the later stages.

The first research corpus, called the Hebrew Objects General Corpus (HOG corpus), with a volume of about 52 thousand words, was constructed from randomly selected contexts, in which O-participant is present in the transitive clause, from the online corpus of Modern Hebrew, Hebrew Web 2021 (heTenTen21) [11]. The main objectives of this corpus were to form a general picture of the use of accusative constructions in Modern Hebrew (by identifying the main types of marked and unmarked objects) and to identify the correlation between the marking strategy and the definite/indefinite status of the referential expression encoding the O-participant of the situation, as well as to record cases when marking within any particular type of referential expression is optional. A total of 1,313 transitive clauses with two participants were found in the HOG corpus.

The second research corpus, the Hebrew Objects Targeted Corpus (HOT corpus), approximately 49 thousand words in volume, included 1,205 two-participant transitive clauses randomly selected from the heTenTen21 corpus. Unlike the HOG corpus, the HOT corpus consisted only of contexts with the types of referential expressions that exhibited the optional object marking in the HOG corpus.

Both research corpora were manually tagged according to 11 parameters. Tagging for parameters associated with the information status of the referent was based on the author's interpretation of the contexts under consideration.

The author’s analysis of statistical data from two research corpora of Modern Hebrew made it possible to identify additional discourse-pragmatic factors that licensed object marking of an asymmetrical type, previously not considered in the scientific literature in relation to Modern Hebrew [24]. Therefore, corpus analysis

was instrumental in carrying out this study in order to identify relevant factors of object marking in Hebrew, based on large quantity of data.

The results of this and similar studies can be used not only for further research, but for both teaching Modern Hebrew and for programming various automated systems that include Hebrew text to improve the accuracy of the information presented. Considering the rapid development of Hebrew-language corpora over the past 20-30 years, as well as the emergence of the rapidly developing technology of artificial intelligence in recent years, we can surmise that corpus-based research of Modern Hebrew will continue to develop, and perhaps the development vector will shift towards combining corpus technologies and AI.

References

1. *Voyeykova M.D.* Vvedeniye. Peterburgskaya shkola funktsional'noy grammatiki: istoriya, sovremennoye sostoyaniye i napravleniya razvitiya [Introduction. St. Petersburg school of functional grammar: history, current state and directions of development] // Acta Linguistica Petropolitana. Trudy instituta lingvisticheskikh issledovaniy. 2015. No.1. P. 7–21.

2. *Lyutikova Ye.A., Tsimmerling A.V., Ron'ko R.V.* Differentsirovannoye markirovaniye argumentov: morfologiya, semantika, sintaksis [Differential marking of arguments: morphology, semantics, syntax] // Voprosy yazykoznaniya. 2016. No. 6. P. 113–127.

3. *Wintner Sh.* Hebrew Computational Linguistics: Past and Future // Artificial Intelligence Review. 2004. No. 21. P. 113–138.

4. *Goldberg Y., Elhadad M.* Word Segmentation, Unknown-word Resolution, and Morphological Agreement in a Hebrew Parsing System // Computational Linguistics, 2013. No. 39(1). P. 121–160.

5. *Yona Sh., Wintner Sh.* A Finite-State Morphological Grammar of Hebrew // Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, Ann Arbor, Michigan. Association for Computational Linguistics. 2005. P. 9–16.

6. *Bar-Haim R., Sima'an K. and Winter Y.* Part-of-speech tagging of Modern Hebrew text // Natural Language Engineering, 2008. No. 14(2). P. 223–251.

7. Hebrew General Corpus [Electronic resource]. URL: https://www.sketchengine.eu/hebrewgc-hebrew-general-corpus/ (date of access: 23.06.2024).

8. CoSIH: The Corpus of Spoken Israeli Hebrew [Electronic resource]. URL: https://cosih.com/ (date of access: 23.06.2024).

9. Map Task Corpus of the Open University of Israel [Electronic resource]. URL: https://www.openu.ac.il/en/academicstudies/matacop/pages/default.aspx (date of access: 23.06.2024).

10. Haifa Corpus of Spoken Hebrew [Electronic resource]. URL: https://cris.haifa.ac.il/en/publications/the-haifa-corpus-of-spoken-hebrew (date of access: 23.06.2024).

11. Hebrew Web Corpus (heTenTen) [Electronic resource]. URL: https://www.sketchengine.eu/hetenten-hebrew-corpus/ (date of access: 23.06.2024).

12. Gutenberg Corpora 2020 [Electronic resource]. URL: https://www.sketchengine .eu/gutenberg-corpora-2020/ (date of access: 23.06.2024).

13. Hebrew Drama Corpus [Electronic resource]. URL: https://www.sketchengine.eu/ dracor-drama-corpora/ (date of access: 23.06.2024).

14. CHILDES Hebrew Corpus [Electronic resource]. URL: https://childes.talkbank.org/ (date of access: 23.06.2024).

15. Hebrew translation corpus [Electronic resource]. URL: https://www.sketchengine.eu /hebrew-translation-corpus/ (date of access: 23.06.2024).

16. *Silber-Varod V., Khorshidi N., Levi L., & Amir N.* The Influence of lexical stress on formant values in spontaneous Hebrew speech. In: The Scottish Consortium for ICPhS 2015 (Ed.), Proceedings of the 19th International Congress of Phonetic Sciences. Melbourne 5–9 August 2019.

17. *Bar-Aba E. B.* Towards a Description of Spoken Hebrew. Hebrew Studies. 2005. № 46. P. 145–167.

18. *Izre'el Sh.* The Basic Unit of Language: A View from Spoken Israeli Hebrew Lecture given at the International Workshop on Afroasiatic Languages, Tsukuba University, March 1-2, 2010 [Electronic resource]. URL: https://www.tau.ac.il/~izreel/ publications/BasicUnitsTsukuba2010.pdf (date of access: 23.06.2024).

19. *Maschler Y.* Discourse markers at frame shifts in Israeli Hebrew talk-in-interaction // Pragmatics. 1997. № 72. P. 183–211.

20. *Maschler Y.* The role of discourse markers in the construction of multivocality in Israeli Hebrew talk-in-interaction // Research on Language and Social Interaction. 2002. № 35. Pp. 1–38.

21. *Gonen E., Zohar L., Noam A.* The Discourse Marker axshav ('now') in Spontaneous Spoken Hebrew: Discursive and Prosodic Features // Journal of Pragmatics. 2015. № 89. P. 69–84.

22. *Dekel N.* A matter of time: tense, mood and aspect in spontaneous Spoken Israeli Hebrew. [Thesis, fully internal, Universiteit van Amsterdam]. 2010. [Electronic resource]. URL: https://pure.uva.nl/ws/files/1041346/77515_thesis.pdf

23. *Weizman E.* Explicitating Irony in a Cross-Cultural Perspective: Discursive Practices in Online Op-eds in French and in Hebrew // Contrastive Pragmatics. 2023. № 4. P. 437–465.

24. *Alekseeva M.Ye.* Odushevlennost' i referentsial'nyy status kak faktory asimmetrichnogo ob"yektnogo markirovaniya v sovremennom ivrite (na primere voprositel'nykh i otnositel'nykh mestoimeniy) [Animacy and referential status as factors of asymmetrical object marking in Modern Hebrew (on the example of interrogative and relative pronouns)]// Litera. 2023. № 6. P. 210–220.