О корпусах современного разговорного иврита

Инна Багратовна Григорян

Российский государственный университет им. А.Н. Косыгина
(Технологии. Дизайн. Искусство)
Москва, Россия

Старший преподаватель
ORCID: 0000-0001-7108-6065
Институт «Академия имени Маймонида»
Кафедра филологии и лингвокультурологии
Российский государственный университет им. А.Н. Косыгина
(Технологии. Дизайн. Искусство)
117997, Россия, Москва, ул. Садовническая, д. 33, стр. 1
Тел.: +7 (925) 099-01-96
E-mail: grigoryan-ib@rguk.ru

DOI: 10.31168/2658-3380.2021.21.4.2

Аннотация: Целью данной работы является охарактеризовать существующие корпусы современного иврита. На данном этапе существует два устных корпуса и один письменный корпус Hebrew Corpus или Лингвистический корпус иврита. Все они составлены профессиональными лингвистами на базе университетов Тель-Авива и Хайфы, а также специалистами из Национального Ближневосточного Языкового Ресурсного центра (National Middle East Resource Centre) Университета Бригама Янга в США. В статье дается подробное описание каждого упомянутого выше корпуса. Также приводятся исследования в области просодии, акустики и других вопросов фонетики; работы по прагматике и дискурсу современного разговорного иврита и более тонким явлениям, связанным, например, с передачей эмоций в речи, которые были выполнены на базе материалов устного корпуса современного израильского разговорного иврита (Corpus of Spoken Israeli Hebrew = CoSIH).

Ключевые слова: *прикладная лингвистика, корпусная лингвистика, корпусы, современный иврит, разговорный иврит, корпус современного разговорного израильского иврита*

На современном этапе развития лингвистики почти ни одно исследование не обходится без привлечения корпусных данных. Корпусную лингвистику относят к разделу прикладной лингвистики, занимающейся «как деятельностью по приложению научных знаний об устройстве и функционировании языка в нелингвистических дисциплинах и в различных сферах практической деятельности человека, а также теоретическое осмысление такой деятельности» [Баранов 2001, стр. 6]. Корпусная лингвистика решает задачи построения лингвистических корпусов данных с использованием компьютерных технологий, а также созданием компьютерных инструментов для их обработки. По одному из определений [Захаров, Богданова 2013, стр. 5], лингвистический корпус – это «большой, представленный в машиночитаемом формате, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач». Другое определение корпуса [Захаров, Богданова 2013, стр. 5], заимствованное из учебника Э. Финегана, определяет корпус как «репрезентативное собрание текстов, включающее информацию о ситуации, …такую как информация о говорящем, авторе, адресате или аудитории». Определений языкового корпуса довольно много, и в каждом из них прослеживаются общие их черты: цель, печатный формат, наличие социолингвистической и металингвистической информации о тексте и репрезентативность данных.

Корпусы делятся на устные (или «речевые»), письменные и смешанные в зависимости от языковых данных, содержащихся в них. Принимая во внимание предназначение корпусов, можно выделить исследовательские корпусы и «иллюстративные», статические и динамические. По другой классификации выделяются многоцелевые и специализированные корпусы. Несмотря на все многообразие формулировок, корпус в первую очередь следует рассматривать как инструмент, как метод работы с собранным материалом либо как базу данных, с помощью которой можно исследовать те или иные языковые явления.

Лингвистический корпус иврита, или hebrewCorpus, был создан в 2009 г. под руководством Шмуэля Болоцкого (Shmuel Bolotzky) по модели arabiCorpus при поддержке National Middle East Language and Resource Center (NMELRC; Университет Бригама Янга, США). Это пример закрытого не пополняемого письменного корпуса на 150 миллионов слов. Корпус содержит тексты разных жанров, включая художественную и научную литературу (журналы), издания СМИ (газеты), статьи из Wikipedia, субтитры к фильмам и религиозные тексты. Данный корпус содержит 23 подкорпуса.

Тексты на иврите тяжело поддаются лемматизации[1] по причине высокой многозначности словоформ: как показали Й. Гольдберг и М. Эльхадад, в современном иврите на словоформу приходится в среднем 2,7 варианта интерпретаций, тогда как в английском – всего 1,4 [Goldberg, Elhadad 2010, стр. 104]. Некоторые словоформы в иврите имеют до 13 различных интерпретаций [Carmel, Maarek 1999, стр. 313]. Поэтому число инструментов лемматизации для иврита невелико. В hebrewCorpus тексты не размечены. Вместо этого программа использует специальные фильтры, которые пытаются предсказать часть речи, основываясь на морфологической структуре. Также используются регулярные выражения, позволяющие значительно улучшить поисковые возможности корпуса.

Несмотря на указанные трудности, можно смело говорить о значительном вкладе специалистов NMELRC в изучение различных лингвистических вопросов, например, таких, как исследование употребления наречия me'ód *'очень'* перед и после прилагательных, сравнительное исследование использования современного иврита в интернет-сообщениях носителей языка и выучивших его в качестве первого иностранного, сопоставительное изучение фразеологизмов с компонентами чувств в современном иврите, английском и русском языках и др. [Руднева 2021, 31–38].

[1] *Лемматизация*, или *морфологическая разметка* текста, — это процедура образования первоначальной формы слова для словоформ в тексте [Захаров, Богданова 2013].

Исторически первым корпусом устной речи стал Лондон-Лунд (The London-Lund Corpus), разрабатывавшийся с 1959 г. Благодаря ему были исследованы использование *actually, really, you know, well* в речи, вопросы и ответы в английском дискурсе, использование пассива, просодических моделей английского разговора. На данном этапе основной интерес исследователей устной речи заключается в изучении способов передачи эмоций, что возможно, например, на базе мультимедийного подкорпуса НКРЯ. В силу ряда причин составление устного корпуса занимает больше времени, чем письменного, а также более трудоемко и материально затратно. При составлении устного корпуса необходима аудио- и видеофиксация речи. Главная трудность в создании фонетических ресурсов подобного рода заключается в необходимости транскрибирования устной речи. Кроме того, возникают проблемы, связанные с выбором алгоритма транскрибирования, особенностями произношения, маркированием нераспознаваемых слов и сопутствующих речи паралингвистических явлений (посторонние шумы).

Начало работ над устным корпусом разговорного израильского иврита планировалось еще на 1998 г. Претворить планы в жизнь удалось спустя 4 года. К этому моменту уже был собран и обработан материал корпуса. Директором программы стал исследователь из Университета Тель-Авива Шломо Изреэл, транскрипцию устной речи в письменную выполнила Нурит Декель. В состав команды, занимающейся разработкой фонетического корпуса, также вошли: Бенджамин Хари (Университет Эмори) – академический менеджер проекта; Джон Дю Буа (Университет Санта-Барбары в Калифорнии), отвечающий за корпусный анализ; Мира Ариэль (Университет Тель-Авива) – дискурсивный анализ и прагматика; Гиора Рахав (Университет Тель-Авива) – социолингвистический анализ и статистика; Эстер Бороховски Бар-Аба (Университет Тель-Авива), исследовавшая синтаксис.

К процессу создания корпуса были подключены десятки волонтеров, которым на начальном этапе объяснили цели исследования, а также взяли письменное согласие на использование полученных в процессе записи данных. Все имена соб-

ственные и другая личная информация в процессе обработки записи были стерты, а на их месте использованы специальные звуки. Материал, полученный в результате многочасовых непрерывных записей, затем компилировался и транскрибировался. При работе над анализом записей была использована программа ELAN, разработанная в Институте психолингвистики Макса Планка (Нидерланды) для аннотации аудио- или видеозаписей с целью их архивирования в рамках программы по сохранению языков, находящихся под угрозой исчезновения. Все тексты сегментированы также по просодическим группам с применением программы Praat, позволяющей анализировать, создавать и управлять звуками.

На подготовительной стадии было оформлено 11 записей, каждая из которых порядка шести часов по продолжительности. На данный момент корпус содержит непрерывные записи повседневной живой израильской речи 53 информантов продолжительностью от 6 до 18 часов. Все записи были произведены в 2001–2002 гг. При отборе материала для публикации использовался культурно-ориентированный подход: информантам предлагалось заполнить анкету, состоящую из вопросов, ответы на которые наиболее полно раскрывали бы их культурный бэкграунд.

Информантам предлагалось заполнить анкету, направленную на сбор социолингвистических данных, образец которой представлен ниже (табл. 1).

Таблица 1. Образец анкеты для социолингвистического контекста

הערות Notes	שאלון Questionnaire	מקום מגורים Place of residence	מצב משפחתי Marital status[2]	מין Sex[3]	השכלה Education[4]
	שאלון[5]	תל-אביב	ר	ז	9

[2] ר=ravak/-a=холост/-ая; נ=nasuy/-a=женат/замужем; ג=garush/grusha=разведенный/разведена.

[3] ז=zakhar=мужской пол; נ=nekeva=женский пол.

[4] Всего лет, потраченных на обучение (включая школьные годы).

[5] http://cosih.com/CoSIH_files/questionnaires/C1_questionnaire.pdf.

גיל Age	מוצא Ethnic origin[6]	שפת אם עברית Native Hebrew speaker[7]	סימון Siglum[8]
22	מ/א	כן	C1

Работа с фонетическими записями, основанная на материале современного разговорного иврита и представленная в итоге в виде речевого корпуса современного иврита, позволила исследовать синтаксис и просодии спонтанной разговорной речи [Silber-Varod 2011], синтаксические особенности высказывания в разговорной речи [Borochovsky Bar-Aba 2010], морфофонологию ивритского корня в современном разговорном языке [Gonen 2008], просодию и акустику современного разговорного иврита [Silber-Varod 2008], а также описать само исследование по корпусной лингвистике в Израиле [Izre'el, Hary and Rahav 2001].

В Университете Хайфы был разработан еще один крупный корпус разговорного иврита, включающий записи общим объемом 17,5 часа следующих жанров: дискурс социального взаимодействия (социальные сети, sms-сообщения, повседневная речь и др.), записи радиопрограмм политического характера, содержащие телефонные разговоры со слушателями. Корпус содержит 572 токена, разметка произведена в основном вручную с редким использованием программы Praat.

Данный корпус закрыт для свободного доступа, так как охраняется авторским правом по решению автора корпуса лингвиста, специалиста по синтаксису разговорного иврита Яэль Машлер (Yael Maschler). На основе данных хайфского корпуса исследовались конструкции с *she…* в предложениях с топикальным выделением в современном иврите [Maschler 2018].

Корпусная лингвистика в Израиле в силу ряда причин находится на начальном этапе развития по сравнению с тем, к

[6] מ=mizrahi=сефард; א=ashkenazi=ашкеназ; ע=aravi=араб.

[7] Этот пункт анкеты предполагает, что информант может и не быть носителем языка.

[8] Условное обозначение для информанта: буквы – C, D, P – по названию учреждений, рекрутировавших информанта; цифра указывает на номер записи и номер диска, с которого была взята запись.

каким результатам пришли в этом направлении лингвисты, работающие, например, на материале английского или русского языков. Тем не менее одной из наиболее удачных попыток создания устного корпуса является CoSIH, или Корпус разговорного израильского иврита, на материале которого были выполнены работы в области морфофонологии, синтаксиса, фонетики, просодии и дискурса лингвистами из университетов Тель-Авива и Хайфы. Помимо устного корпуса разговорного иврита свою лепту в развитие этой области внесли ученые Университета Бригама Янга в США, создавшие Лингвистический корпус иврита, или hebrewCorpus. Несмотря на то, что этот корпус содержит весьма ограниченный материал, он представляет собой инструмент, способный обеспечить исследователя нужной информацией.

Литература

Баранов 2001 – *Баранов А.Н.* Введение в прикладную лингвистику. М.: Эдиториал УРСС, 2001. 360 с.

Захаров, Богданова 2013 – *Захаров В.П., Богданова С.Ю.* Корпусная лингвистика. Учебник для студентов направления «Лингвистика». 2-е изд., перераб. и дополн. СПб.: СПбГУ. РИО. Филологический факультет, 2013. 148 с.

Руднева 2021 – *Руднева О.Р.* Компоненты чувств во фразеологизмах на примерах английского, иврита и русского языков. Выпускная квалификационная работа по спец. 45.03.01 «Филология». М.: ФГБОУ ВО «РГУ им. А.Н. Косыгина», 2021.

Borochovsky Bar-Aba 2010 – *Borochovsky Bar-Aba E.* Ha-'ivrit hameduberet: prakim be-meḥqara, be-taḥbira u-ve-darḵey hav'ata [иврит]. Jerusalem, 2010.

Carmel, Maarek 1999 – *Carmel D., Maarek Y. S.* Morphological Disambiguation for Hebrew Search Systems. Next Generation Information Technologies and Systems. NGITS (1999) / eds. R.Y. Pinter, S. Tsur. Berlin, Heidelberg: Springer, 1999. P. 312–326. (Lecture Notes in Computer Science, vol. 1649). https://doi.org/10.1007/3-540-48521-X_24.

Goldberg, Elhadad 2010 – *Goldberg Y., Elhadad M.* Easy First Dependency Parsing of Modern Hebrew // Proceedings of the North American Chapter of the Association for Computational Linguistics 2010. The First Workshop on Statistical Parsing of Morphologically Rich Languages. Los Angeles, 2010. P. 103–107.

Gonen 2008 – *Gonen I.* Morfologiya šel ha-šoreš be-fo'al be-'ivrit isr'aelit meduberet.'Avodat gmar liqr'at to'ar musmaḵ [иврит]. Tel Aviv: Tel Aviv University, 2008.

Izre'el, Hary and Rahav 2001 – *Izre'el Sh., Hary B. and Rahav G.* Designing CoSIH: The Corpus of Spoken Israeli Hebrew // International Journal of Corpus Linguistics. 2010. Vol. 6. P. 171–197.

Maschler 2018 – *Maschler Y.* The on-line emergence of Hebrew insubordinate *she-*('that/which/who') clauses // Studies in Language. 2018. Vol. 42. Issue 3. P. 669–707.

Silber-Varod 2008 – *Silber-Varod V.* Me'afyaney gvulot šel yeḥidot šel yeḥidot prozodiyot be-'ivrit ha-dvura: nituaḥ tfisati ve-'aqusti. 'Avodat gmar liqrat to'ar musmaḵ [иврит]. Tel Aviv, 2008.

Silber-Varod 2011 – *Silber-Varod V.* The SpeeCHain Perspective: Prosody-Syntax Interface in Spontaneous Spoken Hebrew. PhD dissertation, Tel Aviv University. https://www.openu.ac.il/personal_sites/vered-silber-varod/download/Vered%20Silber-Varod%20Dissertation-7.pdf (дата обращения: 13.12.2021).

Корпусы

https://arabicorpus.byu.edu
https://hebrewcorpus.byu.edu
http://cosih.com/index.html
http://weblx2.haifa.ac.il/~corpus/corpus_website

Hebrew Spoken Corpora

Inna Grigoryan

Russian State University named after A.N. Kosygin (Technologies. Design. Art)
Moscow, Russia

Senior lecturer
ORCID: 0000-0001-7108-6065
Maimonides Academy
Department of philology and linguistic culturology
Russian State University named after A.N. Kosygin (Technologies. Design. Art)
117997, Moscow, Sadovnicheskaya st., 33 build. 1
Tel.: +7 (925) 099-01-96
E-mail: grigoryan-ib@rguk.ru

DOI: 10.31168/2658-3380.2021.21.4.2

Abstract: This article considers to be the characteristics of existed Hebrew corpora. Now there are two spoken corpora and one written, hebrewCorpus or Linguistic Corpus of Hebrew. All these corpora have been made by the professional linguists of University of Tel-Aviv and Haifa and by the language specialists of National Middle East Resource Centre of Brigham Yang in the USA. The article provides full description of each of the mentioned corpus. There is also a list of the research works in morphophonology, syntax, phonetics, prosody, and discourse of Modern Hebrew, both written and spoken which were carried out by the linguists on the basis of the material from CoSIH.

Keywords: *applied linguistics, corpus linguistics, corpora, Modern Hebrew, Modern Colloquial Hebrew, corpus of Spoken Israeli Hebrew*

References

Baranov, A.N., 2001, *Vvedenie v prikladnuiu lingvistiku* [Introduction to Applied Linguistics]. Moscow, Editorial URSS, 360.

Borochovsky Bar-Aba, E., 2010, *Ha-'ivrit ha-meduberet: prakim be-meḥqara, be-taḥbira u-ve-darḳey hav'ata* [Issues on Colloquial Hebrew]. Jerusalem, *Mosad Bialik*. 306.

Carmel D., Maarek, Y.S., 1999, Morphological Disambiguation for Hebrew Search Systems. *Next Generation Information Technologies and Systems. 1999. Lecture Notes in Computer Science*, vol 1649, eds. Pinter R.Y. and Tsur S. Berlin, Heidelberg, *Springer*, 312-326. https://doi.org/10.1007/3-540-48521-X_24.

Finegan, E., 2004 *Language: its structure and use*. NY, *Harcourt Brace College Publishers*.

Goldberg, Y., Elhadad, M., 2010, Easy First Dependency Parsing of Modern Hebrew. *Proceedings of the North American Chapter of the Association for Computational Linguistics 2010. The First Workshop on Statistical Parsing of Morphologically Rich Languages.* Los Angeles, 103–107.

Gonen, I., 2008, *Morfologiya šel ha-šoreš be-fo'al be-'ivrit isr'aelit meduberet.* [Morphophonology of the root in Israeli Spoken Hebrew]. MA in Linguistics, Tel Aviv University.

Izre'el, Sh., Hary, B. and Rahav, G., 2001, Designing CoSIH: The Corpus of Spoken Israeli Hebrew. *International Journal of Corpus Linguistics* 6, 171–197.

Maschler, Y., 2018, The on-line emergence of Hebrew insubordinate she- ('that/which/who') clauses. *Studies in Language* 42:3, 669–707.

Silber-Varod, V., 2008, *Me'afyaney gvulot šel yeḥidot šel yeḥidot prozodiyot be-'ivrit ha-dvura: nituaḥ tfisati ve-'aqusti.* [Phrase termination

characteristics of prosodic items in Colloquial Hebrew: conceptual and acoustic analysis]. MA in Linguistics, Tel Aviv University.

Silber-Varod, V., 2011, *The SpeeCHain Perspective: Prosody-Syntax Interface in Spontaneous Spoken Hebrew.* PhD dissertation, Tel Aviv University. https://www.openu.ac.il/personal_sites/vered-silber-varod/download/Vered%20Silber-Varod%20Dissertation-7.pdf.

Zakharov, V.P., Bogdanova, S.Yu., 2013, *Corpusnaya Lingvistika. Uchebnik dlya studentov napravleniya "Lingvistika"* [Corpus Linguistics. A textbook for students majors in "Linguistics"]. Sankt-Petersburg, SPbGU. RIO. Philologicheskiiy facultet.