A Catalogue of the Hebrew Sounds

Silber-Varod, Vered
veredsv@tauex.tau.ac.il
Tel Aviv University

Cohen, Evyatar
evyatarch@mail.tau.ac.il
Tel Aviv University

Strull, Inbar
inbarstrull@mail.tau.ac.il
Tel Aviv University

Cohen, Evan Gary
evan@tauex.tau.ac.il
Tel Aviv University

Introduction

Our project's goal is to collect the sounds of Hebrew (at utterance, word and phoneme levels) and to make them accessible for linguistic studies and cultural heritage preservation. Our presentation is focused on the pipeline we have developed and on two main tools that we have developed for it: 1. A Hebrew sound parser (The phoneme aligner) and 2. The database platform.

Background

The earliest documentation of Modern Hebrew (henceforth: Hebrew) as a spoken language was in 1916 (Schmelz & Bachi, 1973). Understandably, the first generation of Hebrew speakers could not have acquired an existing form of naturally spoken Hebrew, as there was none at the time (Amir et al., 2019). In fact, the early spoken language developed with only minimal official direction (Gonen, 2019). Hebrew as a native language (i.e., "mother tongue") has been spoken in Israel for approximately a century. During this time, Hebrew has developed and changed, just like any other living language. Communication studies have shown the effectiveness of exposure to authentic audiovisual programs (e.g., news, cartoons, and films) on improving the language proficiency of language learners (Bahrani & Sim, 2012). In Israel, we can list few such influences: Kol Yerushalaim (*The Voice of Jerusalem*), established in 1936, is the earliest public broadcasts of Hebrew and is considered as a turning point from a written language to a spoken language (Liebes & Kampf, 2009). The Instructional Television Centre, launched in 1966, was a ground-breaking effort to use television as a tool to enhance language proficiency (Bahrani & Sim, 2012). Two years later, the Israeli Broadcasting Authority (IBA) launched regular public transmissions.

The collection

Our project is currently in its Proof of Concept stage as we have already developed the pipeline and set of tools that are involved (Figure 1). The current database consists of ~24,000 realizations of phonemes labeled from excerpts of 10 speakers (~1 hour of speech) from the Map Task Corpus recordings (MaTaCOp, Azougi et al., 2015). Our goal is to integrate other audio and video collections from the National Library of Israel and the Israeli Broadcasting Authority (IBA). We will first focus on audio from 1948 to 1968, the 20 years before mass media (mainly IBA) began to influence the speaking style of Hebrew speakers. The ultimate goal is to create a time capsule of how Hebrew sounded over the years, from 1948 until 2022.

The method

The pipeline and set of tools are shown in Figure 1. For large-scale speech processing, we will be using automatic transcriptions of Hebrew, including time stamps and speaker indexing, utilizing one of the few engines available; e.g., Microsoft's Azure Video Analyzer for Media (VAM), an AI tool intended for developer use without requiring machine-learning expertise. We will then add to the orthographic transcriptions at the utterance level, diacritics that represent the vocalization using *nakdanlive* (https://dicta.org.il/), so the Hebrew utterances are represented in vocalized Hebrew script. *Nakdanlive* includes advanced algorithms for predicting the correct pronunciation of homographs. The vocalized script can be readily and accurately converted to standard phonetic transcription (called IPA), using Zemereshet HebrewToIpa tool.

The next two phases in the pipeline are the new innovative tools we have developed (See blue arrows in Figure 1). As soon as we have a full phonetic transcription, we label and align three levels of linguistic units – phoneme, word, and utterance – using a modified EasyAlign (Goldman, 2011), a plug-in for automatic phoneme segmentation. The plug-in runs under PRAAT, a program for acoustic analysis. Our new script that we have developed (https://github.com/evyaco/EasyAlignHebrew) is the first time Hebrew linguists can work with automatic alignment of words and phonemes. We then export each sound unit (audio and label) using the code of EasyAlignHebrew in a structured manner into the Our Voice database into the Hebrew Sounds database <https://ourvoicewebsite.com/SessionsMenu.php>. The database links each sound unit to its context (words and utterances) and to its class of productions in the collections, as well as to the metadata of each speaker and source. Finally, the phonetic search engine enables users to search the database for sequences of phonemes defined by the user and export user-defined collections of sounds.

To conclude, efforts are currently being made to preserve historical and artistic literary texts. However, the voices of the people who revived the language, and even our own voices, those who use the language daily, have not yet been organized and documented in a proper catalogue of phonemes, words, and utterances. It is our mission to bring to the forefront a somewhat neglected cultural heritage unit – the spoken sounds of Hebrew.

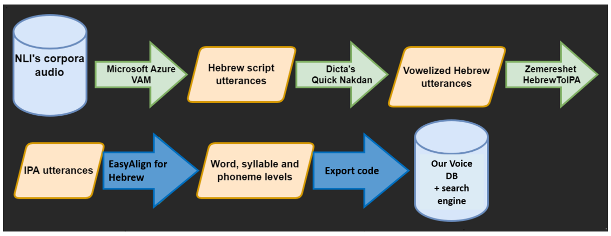

Figure 1: The design of the Hebrew sound extraction and cataloging process

Bibliography

Amir, Noam, Cohen, Evan G. / Reshef, Yael / Gonen, Einat (2019). “Evidence of recent sound change in modern Hebrew – a shift in vowel perception”, in: Calhoun, S. / Escudero, P. / Tabain, M. / Warren, P. (eds.): *Proceedings of the 19th International Congress of Phonetic Sciences,* Canberra, Australia: Australasian Speech Science and Technology Association: 349–352.

Azougi, Jacob / Lerner, Anat / Silber-Varod, Vered (2015-2016). *The Map Task Corpus of the Open University of Israel (MaTaCOp)* . DOI: [04.04.2023]

Bahrani, Taher / Sim, Tam Shu (2012). “Audiovisual News, Cartoons, and Films as Sources of Authentic Language Input and Language Proficiency Enhancement”, in: *Turkish Online Journal of Educational Technology-TOJET* 11, 4: 56–64.

Goldman, Jean-Philippe . (2011). “EasyAlign: an automatic phonetic alignment tool under Praat”, **in:** *Proceedings of InterSpeech* , Firenze, Italy, September 2011.

Gonen, Einat (2019). “Language change, prescriptive language, and spontaneous speech in Modern Hebrew”, in: Doron, Edit / Taube, Moshe / Reshef, Yael / Rappaport Hovav, Malka. *Language Contact, Continuity and Change in the Genesis of Modern Hebrew* , John Benjamins Publishers, 201–220. DOI:

Liebes, Tamar / Kampf, Zohar (2009). “‘Hallo! This is Jerusalem Calling’: The Role of Kol Yerushalayim in the Revival of Spoken Hebrew (1936—1948)”, in: *Cathedra: For the History of Eretz Israel and Its Yishuv* 133: 105–132 (in Hebrew). https://www.jstor.org/stable/23408339 [04.04.2023]

Schmelz, Uziel O. / Bachi, Roberto (1973). Hebrew as the Everyday Language of the Jews in Israel. A Statistical Appraisal. *L#šonénu: A Journal for the Study of the Hebrew Language and Cognate Subjects* 37, 1: 187 – 201 . https://www.jstor.org/stable/2436663