New Tools for Jewish Linguistics

Heidi Lerner

Introduction

For specialized scholars of Jewish linguistics, as well as for general researchers who are fascinated by Jewish languages, online access to the existing and growing network of basic resources that are maximally representative of a particular language or language body is of great value. These resources can range from unanalyzed sound recordings to fully transcribed and annotated text corpora; from dictionaries to the various manifestations of web-based "social media." Even though many of these tools and projects are not yet fully accessible on the Web or remain in various stages of development because of staffing, funding, and technological issues, in the following pages I would like to call attention to their existence and potential benefits. One of the best places to start is the Jewish Language Research Website (jewish-languages.org), which serves as a resource for those studying Jewish linguistics from either an individual or a comparative perspective.

Annotated Corpora

Computer corpora are bodies of computer-readable texts or extracts of written or spoken text that are used for language and linguistic research. Annotated corpora provide scholars with very useful tools for language and linguistic research. Added to the raw text are annotations that describe the linguistic aspects such as morphology, syntax, tone, etc. Benjamin Hary and others have described how Modern Hebrew is underrepresented in corpus linguistics in an article, "Designing CoSIH: The Corpus of Spoken Israeli Hebrew" (*International Journal of Corpus Linguistics:* 6:2 (2002): 171-197). Work is now being done to fill the gaps since the start of the new millennium. The Mila Knowledge Center for Processing Hebrew at the Technion maintains a collection of Modern Hebrew annotated texts at its website (mila.cs.technion.ac.il/english/resources/corpora). These have been organized structurally using Extended Markup Language (XML), a commonly used technology for turning raw or free text into analyzable data, and annotated. Similarly, Tsvi Sadan [also known as Tsuguya Sasaki] of Bar-Ilan University and Jan. H. Kroeze of the University of Pretoria have effectively validated and demonstrated the use of XML as an available tool to transform raw linguistic data into a usable databank for Hebrew linguistic data in their work.

In 1994, Beatrice Santorini of the University of Pennsylvania built a machine-readable parsed and annotated corpus of Yiddish texts (ftp://babel.ling.upenn.edu/research-material/yiddish-corpus). Treebanks are language resources that provide annotations of natural languages at various levels of syntactic structure: at the word level, the phrase level, and the sentence level. The Mila Center has recently released Hebrew Treebank Version 2.0 (www.mila.cs.technion.ac.il/english/resources/corpora/treebank/ver2.0/index.html).

Unannotated Corpora

Unfortunately, carefully annotated corpora are only available for a small number of Jewish languages. Because of copyright issues affecting corpus building, scholars sometimes are forced to turn to machine-readable text collections that are free and open content. Several online text corpora currently are available for Hebrew language research and are still being expanded, such as the Hebrew Wikisource and Eliezer Ben-Yehuda Project. Wikisource is a sister project to Wikipedia that aims to create a free library of primary source texts, and translations of source texts in any language. Hebrew Wikisource (he.wikisource.org) was the first Wikisource non-English language domain. Project Ben-Yehuda's goal (benyehuda.org) is to make freely accessible on the Web the classics of Hebrew literature.

At the recent "2008 Czernowitz Yiddish Language International Centenary Conference" held from August 18-22, 2008 in Czerniivisti, Ukraine, Dr. Cyril Aslanov explored how Wikipedia might be able to provide a window "of visibility" on Yiddish and other such languages.

COMPUTER CORPORA ARE BODIES OF COMPUTER-READABLE TEXTS OR EXTRACTS OF WRITTEN OR SPOKEN TEXT THAT ARE USED FOR LANGUAGE AND LINGUISTIC RESEARCH. ANNOTATED CORPORA PROVIDE SCHOLARS WITH VERY USEFUL TOOLS FOR LANGUAGE AND LINGUISTIC RESEARCH.

Yiddish Wikipedia (yi.wikipedia.org) contains more than five thousand articles, providing access to the usage of Yiddish language in the twentieth century.

Dictionaries
Several Hebrew dictionaries exist on the Web. Maagarim, the Historical Dictionary Project (HDP), is the research arm of the Academy of the Hebrew Language. It aims to "encompass the entire Hebrew lexicon throughout its history"; that is, to present every Hebrew word in its morphological, semantic, and contextual development. This fee-based resource (hebrew-treasures .huji.ac.il) requires registration.

Rav-Milim has been issued by the Melingo Company on the Web in a subscription-based edition (www.melingo.com/rav_ab.htm). The online version offers a variety of features that are not possible in the print version.

The company has also issued Morfix Dictionary, a freely available, online Hebrew-English and English-Hebrew dictionary (milon.morfix.co.il). Morfix is more than just a dictionary or translating tool. It is also an important and effective tool for searching the web. The Morfix Dictionary sits within the Morfix Search Engine, enabling efficient, cross-language morphological searching of websites in Hebrew and English.

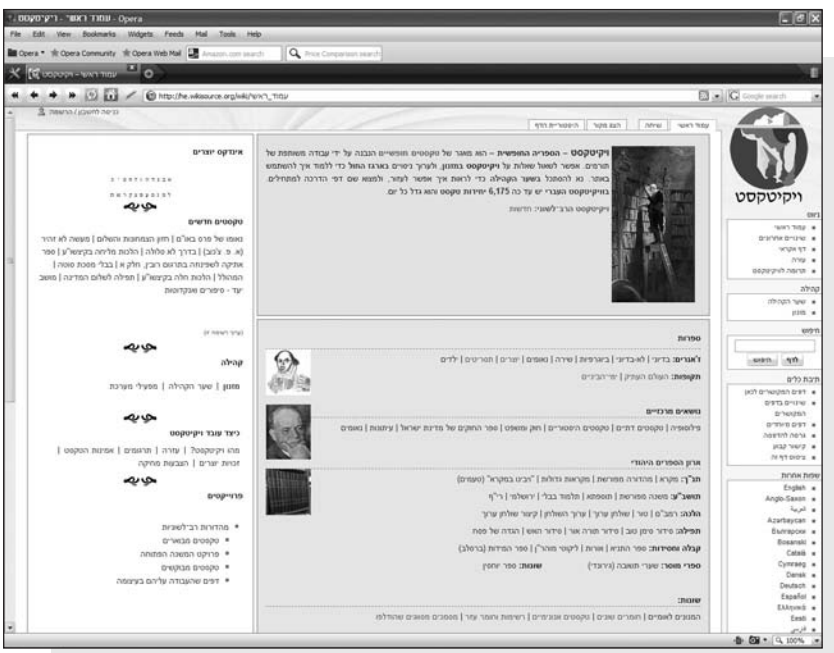
Hebrew Wiktionary (he.wiktionary.org) is part of a multilingual, free dictionary and thesaurus, being written collaboratively by people from around the world. Entries may be edited by anyone.

Yiddish Dictionary Online, (www.yiddishdictionaryonline.com) is a Yiddish-English, English-Yiddish dictionary with English words and phrases and their Yiddish equivalents, with both Hebrew script and romanized spelling, the approximate pronunciation in northern and southern Yiddish, part of speech, and plural versions. It offers word search and alphabetical browsing, rhyming tables, and a few grammatical tables. Authorship of this site cannot be determined and remains unknown.

MUCH HAS BEEN WRITTEN ABOUT THE PROBLEMS OF PROVIDING LONG-TERM PRESERVATION AND ACCESS TO THE ANALOG AND DIGITAL MATERIALS THAT MAKE UP THESE ARCHIVES. AS A FIRST STEP TOWARD MAKING THESE MATERIALS MORE VISIBLE TO THE SCHOLARLY AND OUTSIDE COMMUNITIES, LIBRARIES AND INSTITUTIONS THAT HOUSE THESE RESEARCH COLLECTIONS ARE PUBLISHING THEIR HOLDINGS ON THE INTERNET AND BRINGING VARYING AMOUNTS OF THE COLLECTIONS ONLINE.

The Comprehensive Aramaic Lexicon, hosted by the Hebrew Union College in Cincinnati, aims to create a lexicon of all Aramaic words from 900 BCE up until the early Middle Ages (cal1.cn.huc.edu). The resource consists of a database section with facilities allowing for concordance, dictionary, dialect, and lexicon searches, and a searchable, updated bibliography.

Audio and Sound Collections
The aim of linguistic sound archives is to provide a comprehensive record of the linguistic practices characteristic of a given speech community. Much has been written about the problems of providing long-term preservation and access to the analog and digital materials that make up these archives. As a



Screenshot of the Ben Yehuda Project website, www.benyehuda.org.

first step toward making these materials more visible to the scholarly and outside communities, libraries and institutions that house these research collections are publishing their holdings on the Internet and bringing varying amounts of the collections online. (Note: This article does not include sound archives or repositories that focus on historic recordings of ethnomusicological or liturgical interest.)

The website Eydes: Evidence of Yiddish Documented in European Societies (www.eydes.org/eydes.htm) is devoted to archiving the dialects, folklore, customs, and life experiences of east and central European Jewry. This project is a spinoff of the Language and Cultural Atlas of Ashkenazi Jewry (a decades-long project that was launched at Columbia University by Uriel Weinreich). Within the scope of the project are more than six thousand hours of tape recording taken from 603 separate locales. Also available is an interactive map with audio clips of regional differences in dialect.

Dr. Isabelle Barierre at the Yeled V'Yalda Multilingual Development and Education Research Institute (www.yeled.org/res.asp) has been researching how children develop in different cultural and linguistic settings. Over the past three years she and her team have been recording the interactions of a Yiddish-speaking Hasidic boy with his mother, and hope to publish this corpus soon.

In the 1980s, Dr. Gertrud Reershemius of the University of Aston collected a corpus of spoken Yiddish in Israel. These recordings are now housed at the Phonogrammarchiv, which is part of the Oesterreichische Akademie der Wissenschaften in Vienna (www.pha.oeaw.ac.at). These recordings are slowly being digitized and made available.

SemArch, a project located in the department of Semitic linguistics at the University of Heidelberg, is establishing a digital archive of audio documents (www.semarch.uni-hd.de). Its aim is to archive in digitized form all existing recordings of Semitic dialects and languages and to make them accessible in an Internet database.

Professor Geoffrey Khan of Cambridge University is directing a project that aims to produce a dialect atlas of the surviving North Eastern Neo-Aramaic dialects. It will be a Web-based, free-access catalogue of northeastern Neo-Aramaic languages (Jewish and Christian), searchable by linguistic and grammatical criteria. For the moment, however, researchers can only access an information page (http://nena.ames.cam.ac.uk).

PROFESSOR GEOFFREY KHAN OF CAMBRIDGE UNIVERSITY IS DIRECTING A PROJECT THAT AIMS TO PRODUCE A DIALECT ATLAS OF THE SURVIVING NORTH EASTERN NEO-ARAMAIC DIALECTS. IT WILL BE A WEB-BASED, FREE-ACCESS CATALOGUE OF NORTHEASTERN NEO-ARAMAIC LANGUAGES (JEWISH AND CHRISTIAN), SEARCHABLE BY LINGUISTIC AND GRAMMATICAL CRITERIA.

Members of the staff at the School of Oriental and African Studies, University of London (SOAS) are working with Eli Timan, a native speaker of Iraqi Judeo-Arabic, to document the modern spoken language in the form of audio and video recordings made with speakers in London, Toronto, and Israel. Using ELAN annotation software, Timan has put together a sizeable corpus of partially transcribed recordings, some with time-aligned transcriptions and English translations. Later this year or next, a website will be launched that will have illustrative materials, texts, sound files, images, and possibly some video.

In the public domain, Librivox (librivox.org) provides free audio-books in sixteen languages. The number in Hebrew is still small but growing.

Of the Jewish languages and dialects that have been described and documented, many are now extinct in their spoken form. *The UNESCO Red Book on Endangered Languages: Europe* (www.helsinki.fi/~tasalmin/europe_report.html) and a website produced by Beth Hatefutsoth, the Nahum Goldmann Museum of the Jewish Diaspora, have identified those Jewish languages for which a few speakers remain (www.bh.org.il/links/jewishlangs.asp#Berber). It is incumbent that scholars employ every effort to record and document the last speakers before these languages become fully extinct.

Tools for the Twenty-first Century

Professor Joshua Fishman has noted in an article, "Language Planning for 'The Other Jewish Languages in Israel': An Agenda for the

Beginning of the 21st Century," the dearth of contemporary written texts from Jewish languages such as Judeo-Arabic, Judeo-Persian, and other Jewish languages. Although historic and older texts in these languages exist in libraries and archives around the world, scholars researching them will find little in the way of Web-based or born-digital texts except for those that exist within digitized publications such as dissertations, monographs, and serials. These last resources, which really exist as extensions of print media, have historically been well described, analyzed, and documented by scholars of Jewish languages. To take fullest advantage of the analytical possibilities offered by the computer, an electronic text must first be encoded accurately and consistently, and, even better; include some kind of textual mark-up. Many of the above-mentioned materials cannot be used effectively for computerized linguistic analysis because of problems of transcription and transliteration, and production quality. As the capabilities and quality of optical character resolution (OCR) improve and render these texts machine-readable, scholars of Jewish languages may be able to adapt new methods of linguistic analysis to these bodies of texts.

A project is underway at Université Michel de Montaigne Bordeaux 3 under the direction of Soufiane Rouissi and Ana Stulic to create an electronic edition of a historic Judeo-Spanish text that will serve as a paradigm for corpus building in the context of a collaborative computer-based environment (corpusjudes.p.free.fr/janvier_2006.ppt).

Some linguists are exploring the use of blogs, discussion groups, and other manifestations of Web-based social media as a source of language data. There has been a rapid increase in the number of Yiddish blogs in the past decade. A directory of Yiddish blogs is found at the Tapuz portal (www.tapuz.co.il/forums/main/links.asp?id=516&catId=5300). Ladino is very much alive among members of the online discussion group "Ladinomunita," which has members from all over the world (www.sephardicstudies.org/komunita.html). Also available for the members of this group is a Ladino audio voice chat room on the Internet using the services of Paltalk, the "Salon de Mohabet" as the participants call it.

Researchers are looking at today's use and infusion of Hebrew and Yiddish words into European and Latin American languages. Sarah Benor describes how she has used data from Anglo-Jewish websites such as www.hashkafah.com and www.heebmagazine.com in examining what she refers to "Jewish American English" in her forthcoming article, "Do American Jews Speak a 'Jewish Language'? A Model of Jewish Linguistic Distinctiveness" (*Jewish Quarterly Review*). She has mounted *Jewish English: Distinctive Lexicon* (beta version) on the Jewish Language Research Wiki (sites.google.com/site/jewishlanguageswiki/jewish-english-distinctive-lexicon).

Conclusion

Computerization is playing an increasing role in the study and development of tools and resources for Hebrew and other Jewish languages. Collaborative research and cooperation between individuals, institutions, and government bodies will, in large part, determine how successful and indeed indispensible digital technologies will become for Jewish linguistics. One hopes that these efforts will succeed so that a new generation of tools and applications will soon be readily accessible to all.

Heidi Lerner is the Hebraica/Judaica cataloguer at Stanford University Libraries.