

Resumé „Advancing the potential of Jewish Studies: the Corpus of Modern Yiddish“

Yiddish is an independent fused language which developed in close contact with varieties of (Middle High) German, Hebrew and the Slavonic languages. Once spoken by more than 12 million Jews, Yiddish became an endangered minority language after the Holocaust. Jewish Studies are trying to restore and preserve the remnants of Eastern European Jewish culture, yet they are confronted with several serious problems, such as insufficient availability of texts. The aim of our project is advancing the potential of Jewish Studies by compiling a balanced text corpus of Modern Yiddish of about 10 million words, the first online corpus of Yiddish so far. The Corpus of Modern Yiddish (CMY) is a tool for language documentation, since it contains texts, balanced by sociolinguistic parameters, and provides linguistic annotation of the text (morphological tagging etc.). It enhances the possibilities of research on the Yiddish language as well as on Eastern European Jewish culture, since the Corpus of Modern Yiddish provides convenient data retrieval system for Yiddish texts through a freely accessible web interface. Another aspect to keep in mind is language teaching: in the domain of Yiddish there exists an ever growing number of online courses, which the Corpus of Modern Yiddish complements: CMY is an ideal guide to language usage, and students may refer to the CMY to examine the paradigm of a newly learnt verb, to collect idioms or to learn more about argument or sentence structure. Furthermore, CMY will provide teachers of Yiddish with a virtually inexhaustible source of real-language examples.

The last fifteen years have seen a significant increase of corpus building and corpora-based research. Working with a text corpus provides linguistic research with abundant language data that makes possible even statistical verification of hypotheses. Recently, some languages received so called National Corpora, which serve as a modern powerful tool for various studies and references. Examples are the British National Corpus (100 million words), the Czech National Corpus (500 million words), the Russian National Corpus (150 million words), and the Eastern Armenian National Corpus (90 million words). Since the number and size of Yiddish texts is much smaller than is the case with the languages mentioned above, our aim is to compile a Yiddish corpus of about 10 million words. The CMY is intended to be diversified, i.e. to include texts of various genres (literary and functional texts, such as newspaper articles, operating instructions, commercials etc.), both written and spoken. However, due to the rarity of spoken material and the great efforts (in terms of time and, therefore, money) required to transcribe, written texts will dominate the spoken ones. The structure of the CMY is as follows: two internal subcorpora contain texts from the periods from 1850 to 1939 and from 1940 to the present. We have decided to leave out the period of Early Modern Yiddish between 1740-1850, since the DFG-sponsored project “Historische Syntax des Jiddischen mit transkribiertem Textkorpus zum älteren Jiddisch [Historical Syntax of Yiddish with a transcribed text corpus of Elder Yiddish]” at the Friedrich-Schiller-Universität Jena will offer publicly available texts from this period. The time between 1851-1939 is the blossom of the Yiddish language, whereas the period from 1940 onwards is characterized by the geographical dispersion of the Jews surviving the Holocaust and therefore the rise of varieties of Yiddish, such as Southern and Northern American Yiddish, Israel Yiddish or Eastern European Yiddish.

Since the primary aim of our project is to improve the possibility to conduct research on Yiddish and the culture of its speakers and making the Corpus of Modern Yiddish accessible to as many interested users as possible, the corpus will be available in both Hebrew characters and YIVO-transliteration in Latin characters. For the same reason, the internet site of the corpus will be in both Yiddish and English.

Compiling a corpus requires both linguistic expertise and computer knowledge. The members of the CMY team stand for expertise in linguistics and computation. The Corpus of Modern Yiddish is a joint project of scholars from Regensburg, Germany, and Moscow, Russia.

Prof. Dr. Björn Hansen holds the Chair of Slavic linguistics at the University of Regensburg and leads the DFG project “Korpuslinguistik und diachrone Syntax: die Grammatikalisierung peripherer Subjekte in slavischen Sprachen [Corpus linguistics and diachronic syntax: the grammaticalization of peripheral subjects in the Slavonic languages]”, which provides him with experience in corpus compilation. As a specialist on mood and modality he has conducted research on the modal system of Yiddish together with Sandra Birzer. Sandra Birzer is associate professor of Slavic linguistics at the University of Regensburg. She has taken part in translating Joseph Opatoshu’s novel “A tog in Regensburg” from Yiddish to German. During the last years, the Department of Slavic Studies at the University of Regensburg has developed manifold activities in the field of Jewish Studies, resulting, among others, in a Dilthey Fellowship from the VolkswagenStiftung.

Expertise on Yiddish is provided by Evita Wiecki, lecturer of Yiddish at Regensburg University, and Alexandra Polyak, lecturer of Yiddish at the Chair of Jewish Studies (Moscow State University) and at the Center of Biblical and Jewish Studies (Russian State University for Humanities, Moscow). Among others, she takes part in the project “Charlottengrad und Scheunenviertel. Russische und Osteuropäische Jüdische Migranten im Berlin der 1920-30-er Jahre” (together with O. Budnitsky, G. Pikhan, V. Dohrn, A-K. Sass et alia), which is jointly founded by the Deutsche Forschungsgemeinschaft and the RGNF (Russian National Foundation for the Humanities).

Head of the Moscow-based team members is Prof. Dr. Vladimir Plungian, an internationally renowned language typologist with superior expertise in corpus building, since he is project leader of the Russian National Corpus and the Eastern Armenian National Corpus (EANC).