The London-Lund Corpus 2 of spoken British English (LLC 2)

Carita Paradis, Nele Põldvere & Victoria Johansson
Centre for Languages and Literature, Lund University

Background
With this application, funding was sought for the compilation of a new spoken language corpus, entitled the London-Lund Corpus 2 of spoken British English (LLC 2), designed in accordance to the principles of the London-Lund Corpus (LLC 1), launched in 1975. The goal of compiling a corpus comparable to LLC 1 is to facilitate principled research on recent changes in contemporary spoken English. As opposed to written discourse, spoken discourse has not been documented on a regular basis since the 1990s. This has left a gap in the investigation of naturally occurring language developments in contemporary language use.

The three fundamental stages of corpus compilation for which funding is sought are data collection, transcription and annotation. Like LLC 1, the data for LLC 2 will be recorded at the University College London, and the main emphasis will be put on capturing spontaneous face-to-face conversations between educated adults. Following the data collection process, the recordings will be transcribed and annotated by research assistants. After the completion of these stages, the corpus will be made available to the linguistics community, with the aim of stimulating new and exciting research in the field.

The compilation of LLC 1 was carried out as a collaborative project between Lund University and the University College London. Lund University was mainly responsible for the computerization of the corpus and University College London was the site of recordings of spontaneous face-to-face conversations. More than 50 years have passed since the compilation of LLC 1 started and no similar corpus of both impromptu and scripted speech exists at the moment. This is due to the fact that corpus compilation is a demanding, time-consuming and costly undertaking, and the compilation of a spoken language corpus is considerably more complicated than compiling a corpus based on written sources. LLC 1 was the first spoken language corpus and it was created to provide resources for language research, primarily about the grammar of spoken English. It contains both dialogues and monologues. The dialogue part comprises private conversations, such as spontaneous face-to-face conversations, and public discussions, like broadcast interviews. The monologues are either spontaneous or prepared. For example, spontaneous monologues are represented by running commentaries on sports events or parliamentary debates, while prepared monologues can be lectures or sermons. No comparable spoken language corpora have been compiled after the turn of the century. This has left researchers with an unfortunate gap in the investigation of developments that have taken place over this relatively long period of time.

Purpose
The purpose of this project is to collect contemporary spoken data, to compile a machine-readable corpus of contemporary speech along the lines of the 50 years older LLC 1, to transcribe and annotate the corpus, to make it searchable and easy to use for academics all
over the world. Due to the lack of other spoken language corpora compiled during the past decades, LLC 2 has the potential of attracting a great deal of attention in the linguistics community. The corpus will be made available from the Lund University Humanities Laboratory server that offers secure and long-term storage of digital data. Making LLC 2 accessible to the linguistics community will stimulate research on spoken English in a number of fields in linguistics: prosody, morphosyntax, semantics-pragmatics, information structure, discourse studies, etc. It will allow linguists to make principled comparisons across LLC 1 and LLC 2, to identify and describe the changes that have taken place over the past 50 years, and to make predictions as to where the development of different constructions are heading. Research on language variation and change is crucial for our understanding of the processes that languages are constantly subject to when language users engage in human communication. It is only through real language data that we can learn more about the developments and the dynamic nature of language. The project will be carried out within the research environment of Language, Cognition and Discourse@Lund (http://www.sol.lu.se/en/research/research-groups/797), which is a research group led by myself. The members of the group are interested in how language is used in different situations in speech and writing. LLC 2 will provide the research group with a much-needed and novel dataset for future investigations.

**Method**

The compilation of LLC 2 will be divided into four fundamental stages: data collection, transcribing, markup and annotation, and providing users with free access to the corpus. Funding is sought for the completion of the first three stages. Stage 1 is concerned with the recording of spontaneous face-to-face conversations at the University College London. This requires the purchase of high-quality equipment, which includes recorders, external microphones and mini cameras. Potential participants, both staff and students, will be approached to record speech in different settings, such as common room conversations, seminars, staff meetings, student-supervisor meetings, etc. Also, the recorders will be handed out to participants to capture speech outside of the university, such as dinner table conversations. Stage 2 is the transcribing phase, which concerns the transformation of the recordings into written form. Making transcriptions of spoken language, especially for the purposes of linguistic research, is a time-consuming and demanding job. Therefore, one research assistant will be hired for the year 2016. Assistance is also sought for the completion of Stage 3, i.e. the markup and annotation phase. This entails making the corpus machine-readable and usable for the linguistics community. Therefore, a research assistant, preferably with experience in computational linguistics or computer science, will be hired for the year 2017.