World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:13, No:09, 2019

The Istrian Istrovenetian-Croatian Bilingual Corpus

Authors : Nada Poropat Jeletic, Gordana Hrzica

Abstract : Bilingual conversational corpora represent a meaningful and the most comprehensive data source for investigating the genuine contact phenomena in non-monitored bi-lingual speech productions. They can be particularly useful for bilingual research since some features of bilingual interaction can hardly be accessed with more traditional methodologies (e.g., elicitation tasks). The method of language sampling provides the resources for describing language interaction in a bilingual community and/or in bilingual situations (e.g. code-switching, amount of languages used, number of languages used, etc.). To capture these phenomena in genuine communication situations, such sampling should be as close as possible to spontaneous communication. Bilingual spoken corpus design is methodologically demanding. Therefore this paper aims at describing the methodological challenges that apply to the corpus design of the conversational corpus design of the Istrian Istrovenetian-Croatian Bilingual Corpus. Croatian is the first official language of the Croatian-Italian officially bilingual Istria County, while Istrovenetian is a diatopic subvariety of Venetian, a longlasting lingua franca in the Istrian peninsula, the mother tongue of the members of the Italian National Community in Istria and the primary code of informal everyday communication among the Istrian Italophone population. Within the CLARIN infrastructure, TalkBank is being used, as it provides relevant procedures for designing and analyzing bilingual corpora. Furthermore, it allows public availability allows for easy replication of studies and cumulative progress as a research community builds up around the corpus, while the tools developed within the field of corpus linguistics enable easy retrieval and analysis of information. The method of language sampling employed is kept at the level of spontaneous communication, in order to maximise the naturalness of the collected conversational data. All speakers have provided written informed consent in which they agree to be recorded at a random point within the period of one month after signing the consent. Participants are administered a background questionnaire providing information about the socioeconomic status and the exposure and language usage in the participants social networks. Recording data are being transcribed, phonologically adapted within a standard-sized orthographic form, coded and segmented (speech streams are being segmented into communication units based on syntactic criteria) and are being marked following the CHAT transcription system and its associated CLAN suite of programmes within the TalkBank toolkit. The corpus consists of transcribed sound recordings of 36 bilingual speakers, while the target is to publish the whole corpus by the end of 2020, by sampling spontaneous conversations among approximately 100 speakers from all the bilingual areas of Istria for ensuring representativeness (the participants are being recruited across three generations of native bilingual speakers in all the bilingual areas of the peninsula). Conversational corpora are still rare in TalkBank, so the Corpus will contribute to BilingBank as a highly relevant and scientifically reliable resource for an internationally established and active research community. The impact of the research of communities with societal bilingualism will contribute to the growing body of research on bilingualism and multilingualism, especially regarding topics of language dominance, language attrition and loss, interference and code-switching etc.

Keywords : conversational corpora, bilingual corpora, code-switching, language sampling, corpus design methodology
Conference Title : ICTCLCM 2019 : International Conference on Theoretical Corpus Linguistics and Corpus Methodology
Conference Location : Zurich, Switzerland
Conference Dates : September 16-17, 2019