# Identifying differences between spoken and written language varieties with corpora: What can we learn from EPTIC-SI?

With its multimodal and multilingual design, the EPTIC corpus fosters a range of different research perspectives, involving interpreting and translation and different types of comparisons of the different combinations of subcorpora. EPTIC-SI is, at present, a small multimodal corpus of interpreted and translated Slovene developed within the framework of the EPTIC project and there are plans to further expand it in the future. The initial research on EPTIC-SI has focused on interpreted discourse in contrast with the corresponding translations and the corresponding source texts. In this paper, we aim to expand this research paradigm, by using data from EPTIC-SI and contrasting it with existing monolingual corpora, to shed light on the differences between the spoken and the written varieties of Slovene.

In recent decades, there has been an increasing interest in compiling corpora for Slovene to allow researchers an insight into actual language use. However, as most of the larger existing corpora for Slovene are monolingual, this limits the opportunities to identify or investigate a range of issues arising in multilingual settings, such as translation or interpreting. Moreover, the majority of traditional resources available for Slovene tend to focus mostly on the standard written language, which means that there is much less information available on spoken varieties of the language. This is particularly problematic because, due to historical circumstances, there is a considerable gap between spoken Slovene and the standard written variety in terms of phonology, grammar and discourse. It is therefore not surprising there has been a growing interest in developing resources for the study of spoken Slovene, with GOS, the first corpus of spoken Slovene, as the most comprehensive resource available at present.

While direct comparisons between monolingual corpora of spoken and written language are certainly informative, multilingual comparisons may reveal additional issues less obvious in monolingual contexts. Thus, data from EPTIC-SI can shed light on phenomena emerging in more complex language contact situations. As the interpreted speeches of EPTIC-SI constitute a hybrid form of spoken language, they offer valuable insight into the genres of spoken Slovene when contrasted with GOS, the monolingual spoken corpus comprising several genres, and with the translated speeches of EPTIC-SI, which enable a direct comparison of the same content in spoken and written form. Using the interpreted and translated components of EPTIC-SI and the available monolingual corpora for Slovene, we attempt to explore selected differences arising at discourse level between spoken and written Slovene. Specifically, our analysis focuses on cohesion, as this is one of the discourse features where major differences between written and spoken language can be observed.