The construction and annotation of a spoken corpus for language documentation and research: The Endangered Languages in Contact in Istria and Kvarner project (ELIC)*

Keith Langston
Department of Linguistics
University of Georgia
Athens, GA, USA
langston@uga.edu

Zvjezdana Vrzić
Department of Linguistics
New York University
New York, NY, USA
zv2@nyu.edu

John Hale
Department of Linguistics
University of Georgia
Athens, GA, USA
jthale@uga.edu

Margaret E.L. Renwick
Department of Linguistics
University of Georgia
Athens, GA, USA
mrenwick@uga.edu

***Abstract*— This paper outlines the development of an online spoken corpus aimed at documenting endangered language varieties in Croatia's Istria-Kvarner region. It discusses the challenges and considerations involved in data collection, transcription, and annotation, providing specific examples to elucidate the methodology. The paper also explores the use of existing natural language processing tools for automation, with an aim to offer solutions applicable to other language documentation projects.**

Keywords—Online spoken corpus, endangered languages, Istria-Kvarner region, data collection, transcription, annotation, sociolinguistic interviews, natural language processing, Universal Dependencies, language documentation.

I. INTRODUCTION

This paper describes the construction of an online spoken corpus of endangered language varieties in the Istria-Kvarner region of Croatia, focusing on the decisions that must be made in the collection, transcription, and annotation of data and the challenges posed by creating a corpus of this type. We cite specific examples from data that have been collected and processed to date to illustrate the issues that must be decided and the solutions adopted for this project.

II. DATA AND DATA COLLECTION

The ELIC corpus will include approximately 15 hours of audio data for each of four language varieties spoken in this region, which belong to two different language families: Čakavian Croatian (Slavic) and Istriot, Istro-Romanian, and Istro-Venetian (Romance). These varieties have been in contact with one another for hundreds of years, and beginning in the 20th century have come under increasing pressure from standard Croatian and/or Italian. All of them can be considered endangered and in need of further documentation.

The corpus itself differs from a "prototypical corpus" [1] in a number of ways, which is due to the goals of the corpus and to practical considerations. As a spoken corpus that is intended to document endangered languages, it is relatively small in size and cannot be truly representative or balanced. However, within the limits of the project design, we have tried to include speakers that represent different geographical areas with different local varieties, as well as different age groups. Both for practical reasons and in order to fulfill the research goals of the project, which include the study of language contact phenomena and attitudes towards different varieties, we have opted to collect data in the form of sociolinguistic interviews with individual speakers. While this format arguably does not represent a "natural communicative setting" (per [1], disregarding the problematic nature of defining what is "natural"), the methodology of the sociolinguistic interview [2] is designed to elicit relatively casual/informal speech. In our case, the interview also includes a final set of metalinguistic questions about the use of different language varieties and speaker attitudes towards them. From a practical perspective, an interview of this type generally requires less time to transcribe than a conversation with two or more speakers. However, our interviews are designed to be as conversational as possible, by enlisting speakers of the varieties in question (or of closely related varieties) to conduct the interviews and by not prescribing a rigid format for the interview topics; the interviewers are instructed to follow the lead of the participants and pursue topics that the participant seems more eager to discuss. In several instances, the interviewers already have a relationship with the participant, which also promotes a more conversational atmosphere.

III. TRANSCRIPTION AND ANNOTATION

Our choices for transcription and annotation of the data are also determined by the multifaceted goals of the corpus and the nature of the data themselves. First of all, we expect the data in each interview to be multilingual in nature, because speakers in this region routinely employ two or more varieties in their everyday lives: at a minimum, their local variety and one or both of the standard languages used in the region, Croatian and Italian. One of our goals is to provide a translation into English for all varieties, and also into Croatian for the Romance varieties, to make the data accessible to a wider audience. Since these varieties are not standardized, we must also address the question of how they should be transcribed. For Čakavian, there is an established tradition in scholarly literature of using Croatian orthography, with the addition of a few symbols as needed for sounds that do not occur in standard Croatian. This makes the data easily accessible both to scholars and to members of the local communities who may want to use the corpus for their own purposes (e.g., for language maintenance efforts),

since this also corresponds to longstanding local writing practices. The Romance varieties pose bigger problems, since Istro-Venetian and Istriot speakers use a variety of ad hoc adaptations of Italian orthography, while in the case of Istro-Romanian, there is no established tradition of writing in the local varieties. In these instances, we are creating our own systematic orthographic conventions for transcription, guided by what seems most widespread and acceptable within the communities in question. The use of orthographic transcriptions will still allow for phonetic/phonological analysis, due to the availability of the original audio recordings in the online corpus.

For additional types of annotation we follow standard practices in language documentation, which require translation and grammatical annotation at the word or morpheme level, typically presented in the form of interlinear glosses. We are using a slightly modified version of the Universal Dependencies (UD) tagset [3] for part of speech and morphosyntactic feature labels. In addition to representing an increasingly widely used set of conventions that are intended to be applicable to all languages, the labels used in UD have the advantage of being more transparent than other systems that are widely used in corpora (such as Multext [4] and its adapted version for Central and Eastern European languages, Multext East [5]). Our goal is for the linguistic annotations to be readable by humans, and not just by computers.

The transcription, translation, and other annotations will be time-aligned with the audio at the utterance and word levels. However, even with the relatively limited amount of data to be included in the ELIC corpus, this represents a huge task. For our data, the transcription alone of an hour-long interview takes a minimum of 25 hours and can be much longer, depending on the experience of the transcriber, their familiarity with the specific variety, and the amount of overlapping of the interviewer and speaker. We are facilitating the creation of time-aligned transcriptions by marking the utterance boundaries in Praat TextGrids before they are provided to our research assistants for transcription, but this adds another 5 or more hours to the process. At least some automatic processing of the data is required to make a project of this scope feasible, but unfortunately, dedicated natural language processing (NLP) tools do not exist for these varieties. We are investigating a number of existing tools that can be adapted for our purposes and evaluating their performance. Given the intended uses of our corpus, manual correction will be required to achieve the highest possible level of accuracy, so we must also assess the amount of time required to correct an automated transcription or tagging versus the time required to transcribe and tag the data from scratch. The paper will describe some of the tools that we are currently using and their relative accuracy, as well as plans for future improvements. An additional goal of the ELIC project is to identify solutions that can be easily used by linguists for other language documentation projects.

References

[1] S. Gries and A. Berez, "Linguistic annotation in/for corpus linguistics," in Handbook of Linguistic Annotation, N. Ide and J. Pustejovsky, Eds. Dordrecht: Springer, 2017, pp. 379–409. DOI: 10.1007/978-94-024-0881-2_15

[2] W. Labov, "Field methods of the project on linguistic change and variation," in Language in Use: Readings in Sociolinguistics, J. Baugh and J. Sherzer, Eds. Englewood Cliffs, NJ: Prentice-Hall, 1984, pp. 28–54.

[3] "Morphology: General principles," Universal Dependencies website, https://universaldependencies.org/u/overview/morphology.html (with links to lists of POS and morphosyntactic feature abbreviations), accessed 8 July 2023.

[4] N. Ide and J. Veronis, "MULTEXT: Multilingual text tools and corpora," in COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics, Kyoto, Japan. https://aclanthology.org/C94-1097/, accessed 8 July 2023.

[5] T. Erjavec, "MULTEXT-East," in Handbook of Linguistic Annotation, N. Ide and J. Pustejovsky, Eds. Dordrecht: Springer, 2017, pp. 441–462. DOI: 10.1007/978-94-024-0881-2_17.