Standard annotation for nonstandard speech: tagging the Spisz Dialect Corpus

Helena Grochola-Szczepanek (Institute of the Polish Language, Polish Academy of Sciences), Rafał L. Górski (Institute of the Polish Language, Polish Academy of Sciences), Rupecht von Waldenfels (University of California, Berkeley) and Michał Woźniak (Institute of the Polish Language, Polish Academy of Sciences)

We address the grammatical annotation of a spoken dialectal corpus relying on standard tools. Given the wealth of such tools for Standard Polish, we ask: how can we harness these tools for the annotation of spoken dialectal data?

We seek to annotate data from the Spisz Dialect Corpus which aims at documenting a dialect of southern Poland that exhibits transitional features to Slovak dialects. The corpus is projected to encompass the whole dialect area of Polish Spisz, with multiple speakers of different backgrounds in each settlement to be recorded in open interviews in order to be able to gauge social and areal factors in the dialect and processes of convergence with standard Polish. The interviews are transcribed and made available in an online interface developed in cooperation with similar projects working on Slavic data in other regions.

With the help of this interface, the user is presented with both transcription and the audio in close alignment. For this reason, there is no general need for an exact representation of speech in a phonetic transcription. Rather, it is desirable to normalize and adjust the transcription to Standard Polish as far as possible in order to facilitate different important functions that are trivial to implement with standard data:

- Querying the corpus is much simpler, since the variation inherent in close transcription does not need to be taken in account
- The corpus can be annotated with standard tools that are readily available.

Annotation currently involves morphosyntactic tagging and lemmatization in our corpus; possible future annotation may involve semantic and syntactic annotation as well as close automatic alignment of text segments and speech signal.

In our paper, we describe the approach we have taken to mapping the dialectal phonetic, grammatical and lexical system to the system of the standard variety. While this mapping is in most cases straightforward due to the proximity of the systems, there are a number of issues that are difficult to implement.

We focus on issues of tokenization, the representation of distinctly dialectal phenomena such as lexemes, grammatical morphemes, and syntactic issues such as the markedly different position and form of clitics in this languages, especially the BE-auxiliary marking person/number, which is mostly fused to the verb in standard Polish, but is clearly a second position clitic in this dialect and may, moreover, be replaced by a pronoun in the dialect, positing new problems for morphological annotation. We conclude by evaluating different combinations of mapping strategies and standard tools, hoping that our experience will help other projects facing similar challenges not only in Slavic.