Spisz Dialect Corpus

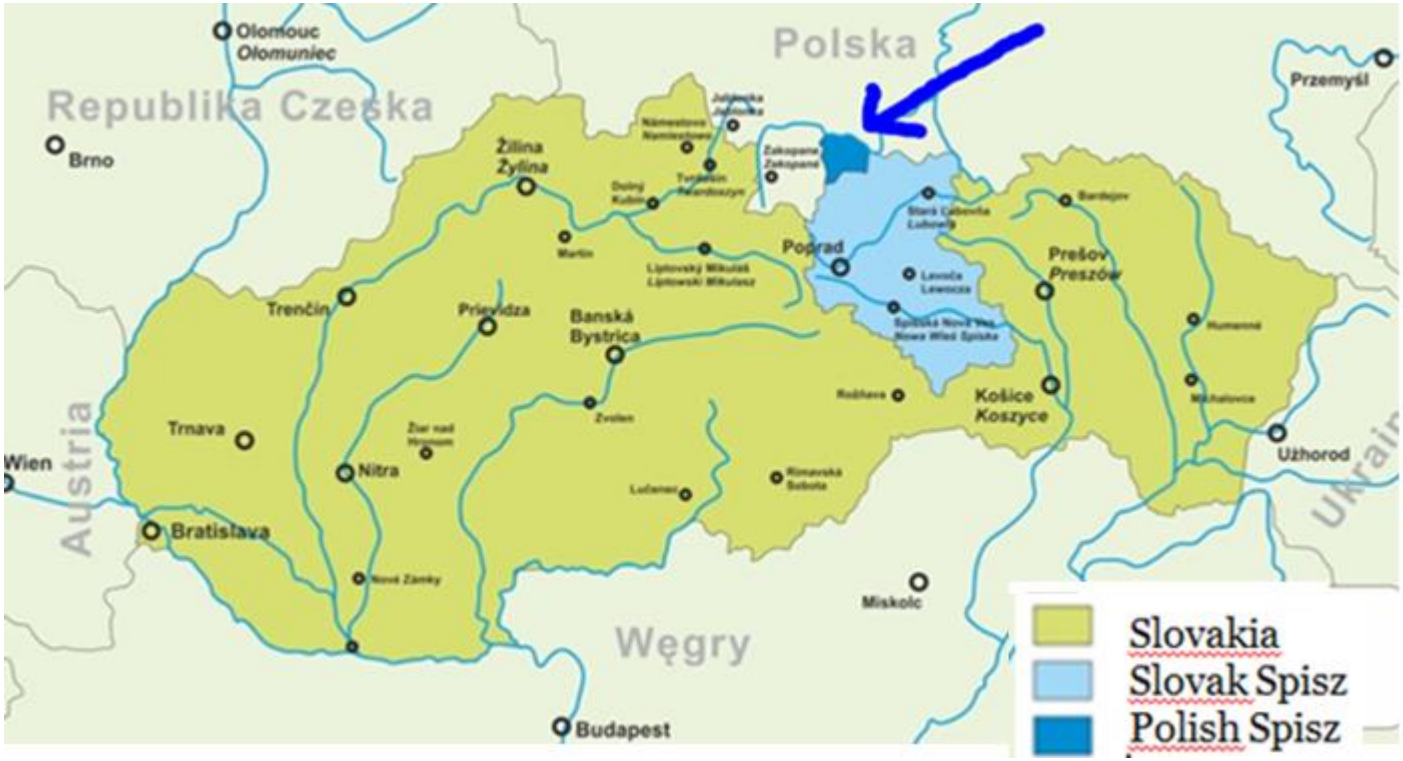
(The Language of Spisz Region.
Corpus of Spoken Texts and Recordings)

https://www.spisz.ijp.pan.pl

The Institute of Polish Language, Polish Academy of Sciences in Krakow
Helena Grochola-Szczepanek
Rafał L. Górski
Ruprecht von Waldenfelds
Michał Woźniak

Spisz Dialect Corpus

It is a collection of texts and recordings documenting the speech of inhabitants of the Polish Spisz region.
Corpus is based on interviews collected in years 2015-2018.



Results

340 informants in 15 villages (corpus includes speech not only older inhabitants, but also the yungest and middle generations, both women and men, people working in various jobs, and also those who feel that they are native to either Poland, Slovakia, or Spisz),

250 hours of recordings (the interviews were recorded in the informants' houses. The files were divided into short segments and saved in the WAV format. We documented diverse and amusing stories concerning the region, its culture, and the fate of its residents and families),

2 mln tokens (the corpus includes forms which are similar to the general Polish variety, but some have a distinct pronunciation, some have unusual morphological patterns, and some have a different meaning. However, the most fascinating group of words includes those characteristic of the region).

Corpus capabilities

Its users gain access to original transcriptions and they have a change to listen to the genuine sound of the Spisz dialect. The corpus is intended for any enthusiast of the regional language and culture. The texts were standardised using the Polish alphabet, hence they can also be read and searched by the users who are not familiar with phonetic alphabets.

The search engine provides its users with various categories of search, for instance: lemmas, tokens, distinct grammatical forms, corresponding segments of a recording, and metadata (by the identification number, age, gender, place of living, education, nationality). The words that are characteristic of the region are explained in a dictionary.

Potential use

This corpus is a new and practical research tool for a modern humanist. The texts are morphologically annotated, which provides scientists with possibilities of conducting studies on inflection, word formation processes, syntax and lexicon. The high-quality recordings are suitable for phonetic studies. The metadata gives opportunities for sociolinguistic researches.

The corpus is also an authentic source of knowledge concerning the rich culture and traditions of the Polish part of Spisz.

We invite you: https://www.spisz.ijp.pan.pl