HELENA GROCHOLA-SZCZEPANEK

KORPUSOWE BADANIA JĘZYKA MIESZKAŃCÓW SPISZA W POLSCE – CELE I ZADANIA

COBISS: 1.01

Korpusne raziskave jezika prebivalcev Spisza na Poljskem – cilji in naloge

V prispevku so predstavljena najpomembnejša vprašanja, povezana z nastajajočim projektom računalniškega korpusa narečnih besedil in posnetkov z območja Spisza na Poljskem. Glavni cilji predstavljenega projekta so: popis govorjenega jezika Spisza, raziskave sodobne jezikovne situacije na tem področju, oblikovanje računalniškega korpusa narečnih besedil in posnetkov ter vzpostavitev dostopa do podatkovne zbirke v obliki spletišča. Avtorica obravnava splošna izhodišča projekta, način pridobivanja gradiva na terenu, vprašanja, povezana s transkribiranjem besedil, in uporabo računalniških orodij.
Ključne besede: raziskave jezika prebivalcev podeželja, korpus spiškega narečja, transkripcija govorjenih besedil

Corpus Research on Language in Spisz, Poland: Objectives and Tasks

This article presents issues related to the newly created electronic corpus of texts and recordings from Poland's Spisz region. The aim of the project is to document spoken language in the Spisz region and to develop an electronic database of dialect texts and recordings. The author discusses the general objectives of the project: obtaining materials during exploratory research, issues related to transcribing texts, and the use of tools.
Keywords: rural dialect research, Spisz dialect corpus, transcription of dialect texts

1 WPROWADZENIE

Badania żywej mowy ludowej należą do jednych z priorytetowych zadań dialektologii. Współczesne procesy globalizacji i unifikacji wpływają na szybko postępujące zmiany w systemach językowych małych społeczności. Duże przeobrażenia dostrzegane są nawet w najbardziej tradycyjnych i podtrzymujących swoją odrębność regionach na południu Polski, np. Podhale, Orawa i Spisz. Odejście dużej części mieszkańców wsi z pracy w rolnictwie i poszukiwanie zatrudnienia w innych miejscach, a także zmiany w strukturze wykształcenia wpływają na otwarcie się społeczeństwa wiejskiego na kulturę i język miasta. Istotny wpływ ma także rozpowszechnianie kultury masowej poprzez

Prispevek je bil predstavljen na 3. Slovenskem dialektološkem posvetu (SDP 3), ki sta ga 11. in 12. februarja 2016 v Ljubljani organizirala Inštitut za slovenski jezik Frana Ramovša ZRC SAZU in Oddelek za slovenistiko Filozofske fakultete Univerze v Ljubljani.

telewizję, prasę, telefonię komórkową i internet. Standardowa polszczyzna używana w szkole, miejscu pracy w mieście, urzędach oraz w telewizji i prasie, jest coraz częściej obecna także w kontaktach językowych mieszkańców wsi. Tradycyjna kultura wiejska uległa procesowi kulturowej interferencji a język ludowy zatracił swoją środowiskową jednolitość na skutek mieszania się środowisk społecznych i wpływu kodu ogólnego.

Współczesne realia na polskiej wsi stawiają przed dialektologią nowe zadania badawcze m.in.: badanie wpływu przemian społecznych i cywilizacyjnych na mowę mieszkańców wsi, zróżnicowanie języka mieszkańców wsi, stopień znajomości i posługiwania się gwarą oraz językiem ogólnym, wariantywność leksykalna, kompetencja językowa i komunikacyjna mieszkańców wsi, świadomość językowa i wartościowanie gwary, procesy integracji i interferencji językowej oraz mechanizmy prowadzące do dezintegracji gwary (por. Pelcowa 2002: 390–391). Dialektolodzy dostrzegają potrzebę prowadzenia aktualnych badań terenowych i rejestrowania współczesnego języka i kultury małych społeczności. Zwraca się uwagę na konieczność nowego podejścia do badań:

> przedmiotem badań musi być już nie tylko gwara, ale dynamiczna i zróżnicowana sytuacja językowa współczesnej wsi, powiązana z opisem różnorodnych czynników ekonomicznych, społecznych, kulturowych i psychologicznych, które mają wpływ na niejednolitość kodu mieszkańców wsi (Grochola-Szczepanek 2013: 49).

Istnieje także potrzeba nowego sposobu opracowania i publikowania rezultatów badań. Współczesne narzędzia informatyczne umożliwiają tworzenie elektronicznych opracowań dialektologicznych, m.in. baz tekstów i nagrań. Zapotrzebowanie na publikacje dostępne online jest ogromne, a korzyści wynikające z tego faktu dla użytkowników są oczywiste (por. Grochola-Szczepanek 2013: 51).

Głównym celem artykułu jest przedstawienie wybranych zagadnień związanych z opracowaniem elektronicznej bazy tekstów i nagrań języka mieszkańców Spisza w Polsce.* Autorka omawia ogólne założenia projektu, cechy przygotowywanej bazy, sposób zbierania materiałów podczas badań terenowych, zagadnienia związane z transkrypcją tekstów oraz wykorzystaniem narzędzi informatycznych.

2 Główne cele i zadania projektu

Głównymi celami prezentowanego projektu są:

- badania współczesnej sytuacji językowej na polskim Spiszu,
- dokumentacja gwary spiskiej,
- opracowanie elektronicznego korpusu tekstów i nagrań gwarowych,
- udostępnienie bazy w formie serwisu internetowego.

* Projekt *Język mieszkańców Spisza: korpus tekstów i nagrań gwarowych* jest finansowany w ramach programu Ministra Nauki i Szkolnictwa Wyższego pod nazwą *Narodowy Program Rozwoju Humanistyki* w latach 2015–2018 (1bH 15 016683).

Kiedy pół wieku temu prowadzono kompleksowe badania językowe na terenie polskiego Spisza (por. Bubak 1972), można je było udostępnić tylko w formie opracowania książkowego. Współczesna wiedza i narzędzia informatyczne pozwalają na opracowanie elektronicznej bazy języka mówionego mieszkańców wsi. Oryginalną mowę mieszkańców wsi będzie można obserwować nie tylko w zapisie, opatrzonym językoznawczymi komentarzami, ale także słyszeć jej naturalne brzmienie. Wyszukiwarka pozwoli na przeszukiwanie tekstów według rozlicznych kryteriów, m.in.: lematów, postaci tekstowej, określonych form gramatycznych, metadanych (np. miejscowość, wiek, płeć, wykształcenie) oraz korespondującego fragmentu nagrania.

Materiały do projektowanej bazy są pozyskiwane poprzez aktualnie i kompleksowo prowadzone badania we wszystkich miejscowościach spiskich.

Zakłada się, że korpus tekstów i nagrań będzie prezentacją języka mówionego mieszkańców Spisza, pozwalającą na poznanie aktualnego stanu gwary oraz jej cech, zjawisk i tendencji. Ważną częścią badań będzie zarejestrowanie i opracowanie autentycznej gwary spiskiej oraz pokazanie zróżnicowania (społecznego i geograficznego) języka mieszkańców wsi spiskich. Jako główną tezę przyjmuje się, że mowa mieszkańców Spisza jest zróżnicowana wewnętrznie przez szereg czynników społecznych. Jednym z najważniejszych faktorów decydujących o stopniu posługiwania się gwarą spiską lub językiem ogólnym jest zróżnicowanie pokoleniowe. Istotne są także inne czynniki społeczne m.in. wykształcenie, zawód, płeć, dłuższe pobyty w mieście i za granicą. Uważa się, że pewne odmienności w mowie występują także pomiędzy poszczególnymi wsiami (por. Bubak 1972; Sowa 1994; Grochola-Szczepanek 2012).

3 OD ZBIORU TEKSTÓW DO KOMPUTEROWEJ BAZY

Badanie mowy mieszkańców wsi i zapisywanie tekstów gwarowych nie jest nowym przedsięwzięciem, ani w polskiej dialektologii, ani w samej gwarze spiskiej. Wystarczy wspomnieć chociażby o zbiorach tekstów gwarowych: Kazimierza Nitscha (Nitsch 1960) oraz Józefa Bubaka (Bubak 1972). Pierwszy zbiór zawiera teksty ze wszystkich dialektów polskich, drugi – tylko z terenu polskiego Spisza. Badania gwary spiskiej przez Bubaka były prowadzone w latach 60. minionego wieku. Jest to najobszerniejszy zapis gwary spiskiej. Autor przeprowadził rozmowy z 38 informatorami. Rozmowy były nagrywane na taśmę magnetofonową, a następnie zapisane pismem fonetycznym. Zbiór liczy około 140 tekstów. Informatorzy zarówno starsi, jak i młodsi posługiwali się jednolitą gwarą, ponieważ społeczeństwo wiejskie było wówczas grupą niezróżnicowaną pod względem społecznym, kulturowym i językowym. Wyraźniej zaznaczyły się jedynie różnice w mowie mieszkańców pochodzących z różnych wsi.

Od kilkudziesięciu lat powstają nowe sposoby opracowywania i prezentowania języka mówionego w postaci elektronicznych baz, głównie w Stanach

Zjednoczonych oraz Europie Zachodniej. W ostatnich latach tego typu projekty są realizowane bądź planowane także w wielu krajach słowiańskich, np. podkorpus gwarowego języka mówionego w *Czeskim Narodowym Korpusie* (*Český národní korpus*), podkorpus dialektów w *Słowackim Narodowym Korpusie* (*Slovenský národný korpus*), podkorpus gwarowy *Narodowego Korpusu Języka Rosyjskiego* (*Национальный корпус русского языка*) (zob. bibliografia).

Warto wspomnieć o pierwszym polskim elektronicznym opracowaniu dialektów polskich autorstwa Haliny Karaś pt. *Dialekty i gwary polskie. Kompendium internetowe* (Karaś red. 2010). W bazie udostępnione są teksty i nagrania z różnych dialektów polskich, w tym także z miejscowości spiskich. W literaturze zostały omówione także prace nad korpusem jednej wsi południowokresowej (Maćkowce), ale publikacja tej bazy nie ukazała się do tej pory (Krawczyk-Wieczorek 2012).

Oblicza się, że w internecie funkcjonuje kilkadziesiąt różnego typu korpusów języka pisanego i mówionego, z czego kilkanaście w krajach słowiańskich (Tkaczewski 2008).

Prezentowany korpus języka mieszkańców Spisza wpisuje się w powszechny obecnie trend prezentowania rezultatów badań w formie elektronicznej. Projekt będzie jednym z pierwszych (o ile w ogóle nie pierwszym) korpusów całościowo przedstawiających język mówiony konkretnego regionu w Polsce on-line.

4 Etapy prac i wykorzystanie/udoskonalenie narzędzi informatycznych

Prace nad projektem można scharakteryzować najogólniej w następujących obszarach:

(1) Badania terenowe – wywiady z mieszkańcami Spisza, rejestrowane przy pomocy dyktafonu cyfrowego oraz archiwizowane w formacie WAV.

(2) Transkrypcja tekstów – zapis w postaci cyfrowej przy zastosowaniu specjalistycznego oprogramowania ELAN, umożliwiającego wygodną anotację materiałów multimedialnych (w szczególności dźwiękowych). Aplikacja ELAN pozwala także na połączenie zapisu audio z jego transkrypcją tekstową, uzupełnienie plików o zestaw metadanych, łatwą edycję, np. wycinanie niepotrzebnych fragmentów, korektę błędów.

(3) Anotacja językoznawcza: materiały uzyskane w transkrypcji będą rozszerzane o warstwę lematyzacji i znakowania morfosyntaktycznego. Do tego zadania użyty będzie tagger języka polskiego (TaKIPI), wykorzystywany przy tworzeniu *Narodowego Korpusu Języka Polskiego* (zob. NKJP, Przepiórkowski i inni 2012). W celu dostosowania go do potrzeb morfologicznych i leksykalnych gwary spiskiej potrzebne będą pewne modyfikacje.

(4) Opracowanie struktury bazy danych korpusu z użyciem programów pozwalających na stworzenie bazy i jej przeszukiwanie (CWB).

(5) Zaprojektowanie i wykonanie serwisu internetowego z dostępnym przez przeglądarkę internetową interfejsem graficznym, umożliwiającym prosty dostęp do korpusu.

5 BADANIA TERENOWE

Badaniami objęty jest cały region Spisza na terenie Polski, w którego skład wchodzi 15 miejscowości spiskich: Czarna Góra, Dursztyn, Falsztyn, Frydman, Jurgów, Kacwin, Krempachy, Łapszanka, Łapsze Niżne, Łapsze Wyżne, Niedzica, Niedzica-Zamek, Nowa Biała, Rzepiska i Trybsz.

Na potrzeby badań terenowych opracowano schemat do wywiadów indywidualnych z mieszkańcami wsi, który składa z 7 części:

(1) Ogólne dane o informatorze (wywiad socjologiczny, metadane).
(2) Wieś, region, mieszkańcy (historia wsi i regionu, pochodzenie mieszkańców, nazwy miejsc, postrzeganie innych).
(3) Szkoła lub praca (tradycyjne prace w gospodarstwie, praca poza gospodarstwem).
(4) Dom, rodzina (małżeństwo, model życia).
(5) Tradycja (zwyczaje, wierzenia, obrzędy, święta, uroczystości rodzinne, wiejskie).
(6) Współczesność, nowoczesność (życie mieszkańców wsi po zmianach, nowe zajęcia, zainteresowania, formy spędzania wolnego czasu).
(7) Kobieta i mężczyzna (różne okresy życia z perspektywy mężczyzny i kobiety, obowiązki męskie i kobiece).

Każda część składa się z szeregu szczegółowych zagadnień i pytań. W zależności od wieku informatora oraz jego doświadczeń życiowych dobierane są odpowiednie zestawy zagadnień i pytań, np. nauka w szkole lub praca w gospodarstwie albo poza gospodarstwem.

Wywiady prowadzone są z osobami w różnym wieku w celu zbadania sytuacji języka mówionego mieszkańców na Spiszu. Istotne w czasie wywiadu jest to, aby rozmówcy posługiwali się swobodnie kodem, którego używają na co dzień. Wypowiedzi starszych i młodszych respondentów pozwolą na obserwowanie zmian językowych w poszczególnych generacjach pokoleniowych. U starszych mieszkańców rejestrujemy głównie autentyczną gwarę spiską. U młodszych zauważa się znajomość kodu gwarowego i języka ogólnego, co w praktyce często realizuje się jako tzw. kod mieszany. Liczba respondentów w poszczególnych wsiach waha od 20 do 35. Uzależnione to jest od ilości oraz jakości pozyskanego materiału nagraniowego. Obliczamy, że z każdej miejscowości zbierze się średnio około 17 godzin nagrań dobrej jakości, co da łącznie ok. 250 godzin nagrań.

Oprócz podstawowej metody badań, jakim jest wywiad indywidualny, sporadycznie prowadzimy także wywiady grupowe. Najczęściej zdarzają się tzw. diady,

czyli rozmowa z dwoma osobami, głównie z małżeństwem. Badania fokusowe można prowadzić także w innych grupach, np. w tzw. rodzinnej grupie fokusowej (rodzice, dzieci, dziadkowie) lub z grupą mężczyzn / kobiet w tym samym wieku lub grupie mieszanej pod względem wieku i płci. Na wsi zdarzają się rodziny wielopokoleniowe (dziadkowie, rodzice i wnuki), mieszkające pod jednym dachem lub w niedużej odległości. Metoda badań fokusowych pozwala na zebranie dużej liczby informacji w stosunkowo krótkim czasie. Materiał zebrany tą metodą jest bogaty i różnorodny. Badacz ma okazję poznać opinie ludzi oraz ich język w bardziej naturalnych dla nich warunkach. Poprzez wypowiedzi kilku uczestników można poznać różne aspekty omawianego problemu, wielowymiarowość zagadnienia (Grochola-Szczepanek 2006). Warto jednak zwrócić uwagę, że w wywiadach grupowych dochodzi często do nakładania się głosów rozmówców, co może wpływać niekorzystnie na czytelność fragmentów nagrania oraz przysporzyć problemów podczas transkrypcji. Czytelny zapis kodu respondenta jest bardzo ważny w opracowaniu bazy tekstów i nagrań. Z tego względu w badaniach korpusowych bardziej stosowną pożądaną metodą pozyskiwania materiałów jest wywiad indywidualny.

Istotnym ogniwem w badaniach terenowych na wsi jest osoba samego badacza. Idealną sytuacją jest, kiedy eksplorator zna kod gwarowy oraz kulturę danego regionu (por. Kąś 2001). Osoba taka wzbudza zaufanie u respondenta, który może swobodnie wypowiadać się w swoim prymarnym kodzie. W przeciwnym wypadku informator może starać się wyeliminować z użycia swój kod gwarowy i próbować przechodzić na kod osoby prowadzącej rozmowę, co wpływa niekorzystnie na wynik badania. Znajomość gwary i kultury danego regionu pomaga w zrozumieniu faktów językowych i kulturowych. W przeciwnym wypadku może powodować nieporozumienia językowe lub nawet niewłaściwą interpretację faktów językowych i kulturowych. Wśród naszych eksploratorów są osoby będące autochtonami, co z pewnością wpływa korzystnie na jakość wywiadów.

Wywiady są rejestrowane przy pomocy dyktafonu cyfrowego oraz archiwizowane na komputerach i dysku zewnętrznym. Pliki zapisywane są w formacie WAV, umożliwiającym bezstratny zapis dźwięku, co jest ważne zarówno ze względu na precyzję transkrypcji, jak i późniejsze wykorzystywanie przez użytkowników korpusu. Konsekwencją tego podejścia jest stosunkowo duży rozmiar plików (godzina nagrań to około 600 MB danych).

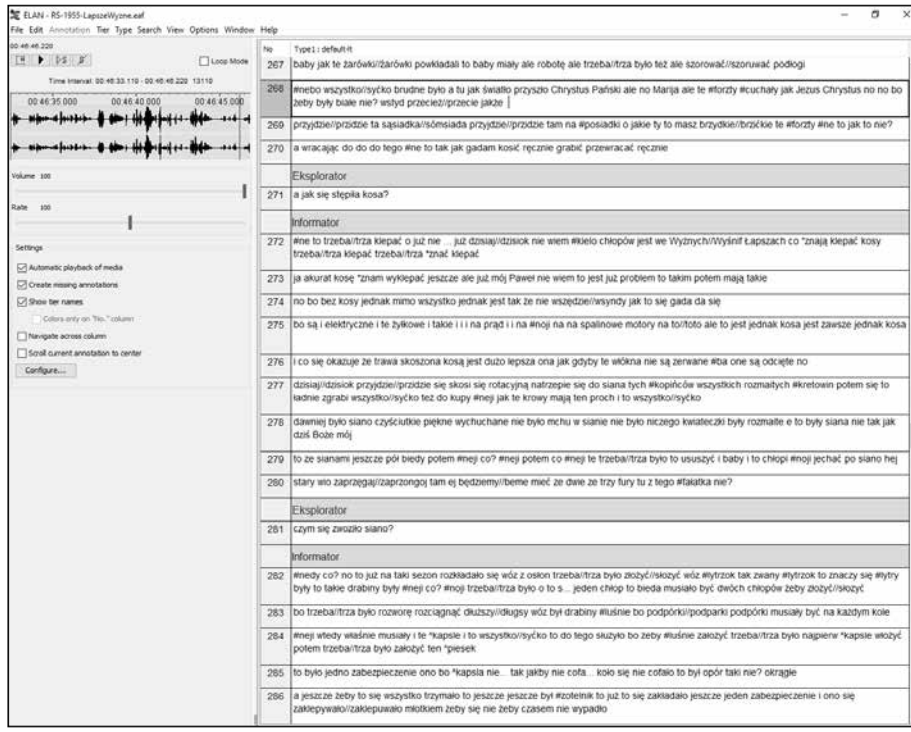
6 Transkrypcja tekstów spiskich

Zapisy materiałów w postaci cyfrowej wykonywane są przy pomocy aplikacji ELAN. Opracowano specjalny szablon uwzględniający warstwy zatytułowane: informator, eksplorator, uwagi, objaśnienia i wątpliwości. Warstwa informatora dotyczy wypowiedzi respondenta. Pytania eksploratora zapisywane są w warstwie: eksplorator. Wszelkie dodatkowe informacje dotyczące nagrania (np. w tle słychać rozmowę domowników) i wypowiedzi informatora (np. śmiech informatora)

Obraz 1: **Transkrypcja tekstów spiskich – ELAN (Annotation Mode)**

zamieszcza się w warstwie: uwagi. W warstwie: objaśnienia notowane są gwarowe wyrazy wraz z objaśnieniami. Wszelkie wątpliwości dotyczące zapisu, formy lub objaśnienia jakiegoś wyrazu sygnalizuje się w ostatniej warstwie: wątpliwości. Program pozwala na swobodne przemieszczanie się po poszczególnych segmentach warstw, wyszukiwanie dowolnego segmentu zapisu w połączeniu z odpowiednim fragmentem nagrania, różne tryby przeglądania transkrypcji i nagrania.

W transkrypcji tekstów spiskich stosuje się zapis ortograficzny, ponieważ tylko transkrypcja ogólna pozwala na wykorzystanie istniejących narzędzi do lematyzacji i anotacji morfosyntaktycznej, zaprojektowanych na potrzeby polszczyzny ogólnej. Pisownia ortografizowana lub obocznie fonetyczna / półfonetyczna / ortograficzna jest powszechnie stosowana w wielu opracowaniach korpusowych, por. *Korpusy mluvené češtiny ORAL2006*, *ORAL2008*, *SCHOLA2010*, *BMK*, *PMK*, *GOS: referenčni govorni korpus slovenskega jezika*, *Национальный корпус русского языка*, *Freiburg English Dialect Corpus*, *Nordic Dialect Corpus* (zob. bibliografia). Zapis ortograficzny jest z powodzeniem stosowany w korpusowym opracowaniu gwary jednego z regionów Rosji *The Ustya River Basin Corpus*, co zostało przedstawione w artykule: *Why Standard Orthography?: Building the Ustya River Basin Corpus* (Waldenfels – Daniel – Dobrushina 2014).



Obraz 2: Transkrypcja tekstów spiskich – ELAN (Transcription Mode)

W trakcie opracowywania transkrypcji tekstów uwzględnia się 2 poziomy anotacji: ogólny i gwarowy. Ogólny powstaje przede wszystkim na potrzeby taggera i sprowadza wszelkie standardowe zmiany fonetyczne (np. mazurzenie, samogłoski pochylone, rozkład nosówek) do postaci ogólnej. Poziom gwarowy uwzględnia oprócz oryginalnych typowo gwarowych form także zmiany morfologiczne, np. inny morfem, inny paradygmat fleksyjny, aglutynat dołączony do innej części mowy lub wolnostojący. Anotacja gwarowa zapisywana jest przy użyciu znaków ortografii ogólnej, ale z zachowaniem cech wymowy gwary spiskiej. Poziom anotacji ogólnej jest tworzony sztucznie, natomiast poziom notacji gwarowej jest bliższy rzeczywistości.

Podczas transkrypcji oba poziomy zapisywane w jednej warstwie. W wersji końcowej oba poziomy (ogólny i gwarowy) znajdą się w dwóch różnych warstwach. Do notowania informacji o określonych typach odmienności w kodzie gwarowym niezbędny jest system umownych symboli stosowanych w transkrypcji, m.in.:

// inna postać wyrazu: *mebli//meblów*, *pragnęła//pragła*, *widzisz//wis*,
. aglutynat dołączony do innej części mowy: *portki.f*, *teroz.ek*, *wnet.ef*,
wyraz dyferencyjny (unikalny): *#ino*, *#samopiyrse*, *#wakuwali*,
^ wyraz dyferencyjny (homonimiczny z językiem ogólnym), np. *^boisko*, *^dziadowina*, *^przodek,*

*	związki wyrazowe, składniowe, np. *młoda pani*, *rada widzieć*, *znać wyklepać*,
[?]	fragment niepewny, np. *te tereny dużo zalesiali tam pocz... [?] zresztą to największe tereny jak chodzi o las w Polsce*,
<…>	wtrącenia z innych języków, gwar, np. *ja ja spał w <Gartenhaus-ie> elegancki domek miałem//miołef*,
…	wyraz lub wypowiedź urwana, np. *dosta... dostawali możliwość wyjazdu.*

Przykład fragmentu zapisu:

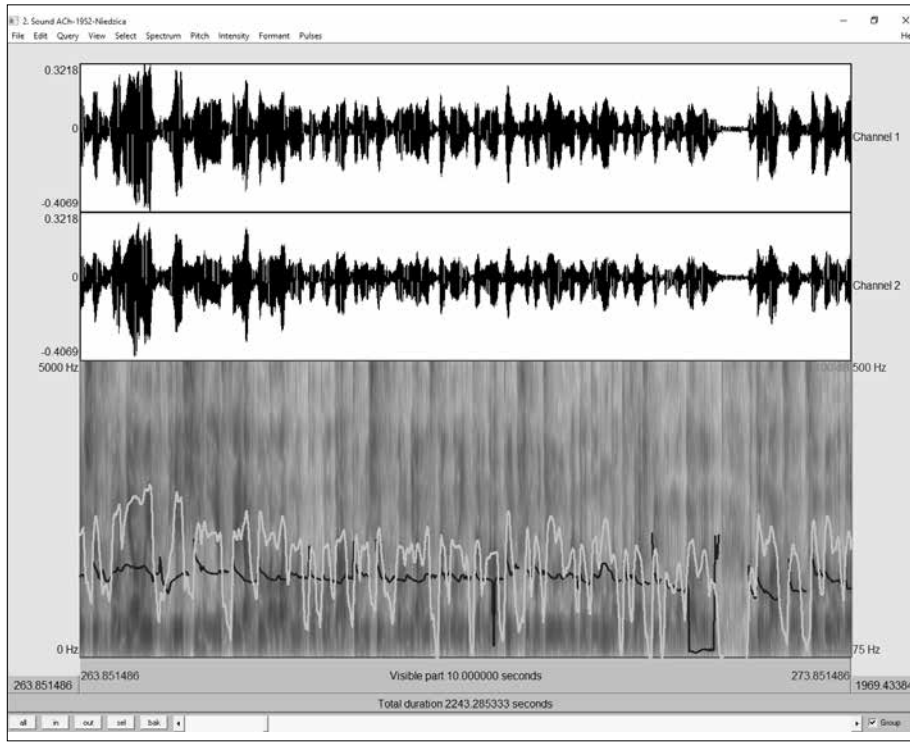
Wersja oryginalna, czyli tak, jak mówi informator	**Transkrypcja: poziom ogólny i gwarowy łącznie**
zamiast nafta to sie wołało gaz o gaz kup mi tam Angieła gazu bo ni ma w lampie /	*zamiast nafta to się ^wołało ^gaz o ^gaz kup mi tam Aniela//Angieła ^gazu bo nie//ni ma w lampie /*
noji lampy próndu nie było jesce jo pamiyntóm taki cas ze bo to za mnie tak prónd zakładali to próndu nie było to Jezus kochany /	*#noji lampy prądu nie było jeszcze ja pamiętam taki czas że bo to *za mnie tak prąd zakładali to prądu nie było to Jezus kochany /*
baby jak te żarówki powkładali to baby miały ale robote ale trza było tyz ale szoruwać podłogi /	*baby jak te żarówki//żarówki powkładali to baby miały ale robotę ale trzeba//trza było też ale szorować//szoruwać podłogi /*
ne bo syćko brudne było a tu jak światło przisło Krystus Pański ale no Marija ale te forzty cuchały jak Jezus Krystus no no bo zeby były biołe nie? /	*#ne bo wszystko//syćko brudne było a tu jak światło przyszło Chrystus Pański ale no Maria// Marija ale te #forzty #cuchały jak Jezus Chrystus no no bo żeby były białe nie?*
wstyd przecie jakze przidzie ta sómsiada przidzie tam na posiadki o jakie ty to mos brzićkie te forzty ne to jak to nie?	*wstyd przecież//przecie jakże / przyjdzie// przidzie ta sąsiadka//sómsiada przyjdzie// przidzie tam na #posiadki o jakie ty to masz brzydkie//brzićkie te #forzty #ne to jak to nie?*

Każdy wyraz zaznaczony w anotacji jednym ze znaków # ˆ * jest objaśniony w warstwie: objaśnienia. Wyrazy z przykładowego powyższego zapisu objaśniono następująco:

cuchać czas. ndk ‘szorować’,
forzty rzecz. nmos ‘podłoga z desek’,
gaz rzecz. mnż ‘tu: nafta’,
ne ‘połączenie mod. no i spój. a, będące wynikiem szybkiego tempa mowy’,
noji ‘połączenie mod. no i spój. i, będące wynikiem szybkiego mówienia’,
posiadki rzecz nmos blp ‘towarzyskie odwiedziny u kogoś’,
wołać czas. ndk ‘tu: nazywać kogoś lub coś’.

Wyrazy ze znakiem // nie są objaśniane, gdyż mają to samo znaczenie co formy ogólne, do których zostały sprowadzone, np. *przecież//przecie*, *szorować//szoruwać*, *trzeba//trza*, *żarówki//żarówki*, *sąsiadka//sómsiada*, *wszystko// syćko*.

Najlepszą dokumentacją autentycznej gwary mieszkańców Spisza są nagrania w oryginale. Udostępniony katalog plików dźwiękowych w bazie pozwoli na



Obraz 3: **Analiza fonetyczna – PRAAT**

dokładne badania zjawisk fonetycznych w późniejszym czasie , np. przy pomocy takich programów, jak PRAAT.

7 Podsumowanie

Na zakończenie warto zwrócić uwagę na dwie istotne cechy projektowanego opracowania: dokumentacja języka i kultury mieszkańców Spisza oraz nowatorski sposób gromadzenia danych, który daje m.in. możliwość przeszukiwania tekstów i nagrań, połączenie zapisu i dźwięku oraz edycji i poszerzania bazy w przyszłości.

Badania języka i kultury małych grup społecznych poszerzają i wzbogacają wiedzę na temat kultury narodowej. Prace badawcze nad odmiennością wielokulturowego i wieloetnicznego regionu mają szczególne znaczenie w dobie zacierania się różnic i zmierzania w kierunku kultury masowej. Od badań Józefa Bubaka, w ciągu półwiecza język mieszkańców Spisza wyraźnie ewoluował. Aktualne badania na tym terenie poszerzą wiedzę na temat gwary spiskiej, wskażą kierunek zmian językowych oraz współczesne tendencje. Teksty dostarczą także wiedzy na temat kultury materialnej i niematerialnej, życia społecznego,

zwyczajów, obrzędów społeczności wiejskiej na Spiszu. Korpus tekstów i nagrań z polskiego Spisza będzie dokumentacją języka mieszkańców wsi w jego naturalnym, wiejskim środowisku na początku XXI wieku.

Elektroniczna baza jest najbardziej nowatorskim sposobem gromadzenia danych i tworzenia kompendium internetowego w obecnym stanie rozwoju lingwistyki komputerowej. Zarejestrowany i opracowany język tradycyjnej społeczności spiskiej, przedstawiony w formie nowoczesnej bazy danych pozwoli na dokonywanie różnego typu kwerend prostych i złożonych. Opracowanie będzie dedykowane nie tylko wąskiemu gronu specjalistów znających pismo fonetyczne, lecz w zasadzie każdemu użytkownikowi. Baza umożliwi swobodny dostęp do nagrań, czyli oryginalnego materiału, a nie tylko jego translacji na tekst pisany. Kiedy pół wieku temu robiono wywiady na polskim Spiszu, można było je udostępnić tylko w formie pisemnej. Elektroniczna forma edycji daje możliwość dokonywania zmian oraz poszerzania bazy o dalsze badania w późniejszym czasie.

LITERATURA

Bubak 1972 = Józef Bubak, Spiskie teksty gwarowe z obszaru Polski, *Zeszyty Naukowe UJ*, Kraków, 1972 (Prace Językoznawcze 36).

Český národní korpus = *Český národní korpus*, http://ucnk.ff.cuni.cz/, https://trnka.korpus.cz/~lukes/files/LREC_A0+.pdf [27.08.2016].

Freiburg English Dialect Corpus = *Freiburg English Dialect Corpus*, http://www2.anglistik.uni-freiburg.de/institut/lskortmann/FRED/index.htm [27.08.2016].

GOS = *Referenčni govorni korpus slovenskega jezika* (GOS), www.korpus-gos.net [27.08.2016].

Grochola-Szczepanek 2006 = Helena Grochola-Szczepanek, Badania fokusowe mowy mieszkańców wsi, *Socjolingwistyka* 20 (2006), 19−35.

Grochola-Szczepanek 2012 = Helena Grochola-Szczepanek, *Język mieszkańców Spisza: płeć jako czynnik różnicujący*, Kraków: IJP PAN, 2012 (Prace IJP PAN 139).

Grochola-Szczepanek 2013 = Helena Grochola-Szczepanek, Badania języka mieszkańców wsi w kontekście przemian społecznych, *Socjolingwistyka* 27 (2013), 43–53.

Karaś 2010 = Halina Karaś (red.), *Gwary polskie. Kompendium internetowe*, Zakład Historii Języka Polskiego i Dialektologii UW, Towarzystwo Kultury Języka, 2010, www.gwarypolskie.uw.edu.pl [27.08.2016].

Kąś 2001 = Józef Kąś, Metodologia badań leksyki gwarowej w kontekście współczesnych przemian kulturowych i społecznych, w: *Gwary dziś* 1: *Metodologia badań*, red. Jerzy Sierociuk, Poznań: Łódzkie Towarzystwo Naukowe, 2001, 191–200.

Korpusy mluvené češtiny = *Korpusy mluvené češtiny ORAL2006, ORAL2008, SCHOLA2010, BMK, PMK*, http://ucnk.ff.cuni.cz/struktura.php [27.08.2016].

Krawczyk-Wieczorek 2012 = Aleksandra Krawczyk-Wieczorek, Automatyczna lematyzacja tekstu w zapisie fonetycznym: korpus polskiej gwary południowokresowej, *Język polski* 92 (2012), nr 1, 11–19.

NKJP = *Narodowy Korpus Języka Polskiego*, http://nkjp.pl [27.08.2016].

Nitsch 1960 = Kazimierz Nitsch, *Wybór polskich tekstów gwarowych*, Warszawa: Państwowe Wydawnictwo Naukowe, ²1960.

Nordic Dialect Corpus and Database = *Nordic Dialect Corpus and Database*, http://www.tekstlab.uio.no/nota/scandiasyn/ [27.08.2016].

Pelcowa 2002 = Halina Pelcowa, Dialektologia wobec wyzwań XXI wieku, w: *Dialektologia jako dziedzina językoznawstwa i przedmiot dydaktyki: księga jubileuszowa dedykowana Profesorowi Karolowi Dejnie*, red. Sławomir Gala, Łódź: Łódzkie Towarzystwo Naukowe, 2002, 383−392.

Przepiórkowski i inni 2012 = Adam Przepiórkowski i inni (red.), *Narodowy Korpus Języka Polskiego*, Warszawa: Wydawnictwo Naukowe PWN, 2012, http://nkjp.pl/index.php?page=3&lang=0 [27.08.2016].

Slovenský národný korpus = *Slovenský národný korpus*, http://korpus.sk/dialect.html [27.08.2016].

Sowa 1994 = Franciszek Sowa, *System fonologiczny polskich gwar spiskich*, Wrocław – Warszawa – Kraków: Ossolineum, 1994.

Tkaczewski 2008 = Dariusz Tkaczewski, Český národní korpus – internetowe źródło standaryzacji i weryfikacji języka czeskiego oraz nowoczesne narzędzie dydaktyczne, *Bohemistyka* 8 (2008), nr 1–4, 363–378, http://www.bohemistyka.pl/artykuly/2008/ART_Tkaczewski.pdf [27.08.2016].

Waldenfelds – Daniel – Dobrushina 2014 = Ruprecht von Waldenfels – Misha Daniel – Nina Dobrushina, Why Standard Orthography?: Building the Ustya River Basin Corpus, an Online Corpus of a Russian Dialect, *Dialog* 21 (2014), http://www.dialog-21.ru/digests/dialog2014/materials/pdf/WaldenfelsR.pdf [27.08.2016].

Национальный корпус русского языка = *Национальный корпус русского языка*, www.ruscorpora.ru/search-dialect.html [27.08.2016].

Povzetek

Korpusne raziskave jezika prebivalcev Spisza na Poljskem – cilji in naloge

V prispevku so predstavljena najpomembnejša vprašanja, povezana z nastajajočim projektom računalniškega korpusa narečnih besedil in posnetkov z območja Spisza na Poljskem.

Glavni cilji predstavljenega projekta so: popis govorjenega jezika Spisza, raziskave sodobne jezikovne situacije na tem področju, oblikovanje računalniškega korpusa narečnih besedil in posnetkov ter vzpostavitev dostopa do podatkovne zbirke v obliki spletišča. Avtorica obravnava splošna izhodišča projekta, način pridobivanja gradiva na terenu, vprašanja, povezana s transkribiranjem besedil, in uporabo računalniških orodij.

Gradivo za načrtovano zbirko podatkov se trenutno zbira na kompleksno zastavljenih raziskavah v vseh spiških krajih. Za transkribiranje besedil v elektronski obliki je bilo uporabljeno specializirano orodje ELAN, ki omogoča označevanje multimedijskega gradiva (še posebej zvočnega). Na naslednji stopnji bo pridobljeno gradivo razširjeno s plastjo jezikoslovnih oznak: lematizacijo ter oblikoslovnim in skladenjskim označevanjem. Končni dosežek predstavljenega projekta bo računalniški korpus narečnih besedil in posnetkov s področja Spisza, ki bo dostopen na medmrežju, s prijaznim grafičnim vmesnikom, z orodji za enostavno in zahtevno iskanje ter s slovarjem spiškega narečja.

Nove raziskave območja Spisza so v obdobju zabrisovanja tradicionalnih razlik in premika v smeri množične kulture izredno pomembne. Od zadnjih raziskav na tem področju, ki so potekale pred pol stoletja, se je jezik prebivalcev vasi izrazito spremenil. Trenutne raziskave bodo razširile vednost o spiškem narečju in pokazale smer jezikovnih sprememb oz. sodobne razvojne težnje.