FOLIA 332

Annales Universitatis Paedagogicae Cracoviensis

Studia Linguistica 16 (2021)
ISSN 2083-1765
DOI 10.24917/20831765.16.5

Helena Grochola-Szczepanek
ORCID 0000-0002-1511-0486
Instytut Języka Polskiego PAN, Kraków, Polska

Od nagrania do korpusu, czyli o metodzie archiwizowania języka mówionego mieszkańców wsi z wykorzystaniem narzędzi lingwistyki cyfrowej

Słowa kluczowe: korpus, język mówiony, gwara, transkrypcja

Keywords: corpus, spoken language, dialect, transcription

Wprowadzenie

Język mówiony jest genetycznie pierwotną i pierwszą reprezentacją języka naturalnego w odniesieniu do pisma, które pełni rolę wtórną (por. Labocha 2012: 139). Dialogowy charakter, ulotność, potoczność oraz emocjonalność mowy sprawiają, że archiwizowanie wypowiedzi ustnych nie należy do łatwych przedsięwzięć. W czasach, kiedy nie było jeszcze możliwości nagrywania, jedynym potencjalnym sposobem utrwalania mowy była pamięć ludzka, a następnie pismo. Dokumentaliści byli w stanie zanotować ze słuchu jedynie fragmenty języka mówionego, np. określone nazwy, formy, wyrażenia. Nie można było utrwalić dokładnie całej wypowiedzi. Eksploratorzy podczas badań terenowych nie nadążali z zapisywaniem tekstu rozmowy. Musieli prosić wielokrotnie o powtórzenie jakiegoś fragmentu przez informatora, co ostatecznie zaburzało naturalny rytm wypowiedzi (por. Lewaszkiewicz 2017: 186). Luki powstałe w wypowiedzi uzupełniane były przez samego badacza w celu utrzymania spójności tekstu. Dość powszechna praktyka poprawiania tekstów utrudniała w rezultacie prowadzenie analiz składniowych (por. Dunaj 1986: 16). Pierwsze magnetofony w latach 50. i 60. ubiegłego wieku tylko w niewielkim stopniu wpłynęły na możliwość rejestrowania mowy. Był to sprzęt ciężki i niewygodny do obsługi w terenie, a nośniki do nagrywania słabej jakości oraz małej pojemności (por. Sierociuk 2009: 182–183). Prawdziwy postęp w nagrywaniu języka mówionego przyniosły dopiero dyktafony, które dzięki małemu rozmiarowi i wysokiej jakości nagrania świetnie nadają się do rejestracji mowy podczas badań terenowych. Dłuższe nagrywanie pozwala na uzyskanie swobodniejszych, naturalnych wypowiedzi, co z kolei jest podstawą do badań nad mówioną odmianą języka. Istotną kwestią w najnowszych metodach archiwizacji języka mówionego jest wykorzystanie narzędzi informatycznych do przetwarzania mowy oraz tworzenia elektronicznych

korpusów z warstwą dźwiękową. Samo przetwarzane mowy na tekst pisany jest dosyć dużym postępem, jednak automatyczny wynik zawiera liczne błędy i musi być poddany skrupulatnej ręcznej adiustacji. Ponadto zaznaczyć trzeba, że programy do przetwarzania obejmują tylko odmianę ogólną języka, nie sprawdzają się natomiast przy odmianach gwarowych i regionalnych. Narzędzia do przetwarzania konkretnego systemu gwarowego to dopiero przyszłość lingwistyki komputerowej. Na razie pozostaje ręczna transkrypcja nagrań gwarowych. Archiwizowanie danych mówionych, czy to ze standardowej odmiany języka, czy regionalnej i gwarowej, w postaci elektronicznego korpusu językowego jest obecnie najbardziej zaawansowaną i praktyczną metodą gromadzenia zasobów mowy oraz udostępniania jej do badań naukowych. Język mówiony w odróżnieniu od pisanego stwarza znacznie więcej problemów podczas dokumentowania, ale w rezultacie nowoczesny korpus elektroniczny daje szerokie możliwości badań nad żywą mową.

Głównym zamierzeniem niniejszego tekstu jest omówienie wybranych zagadnień związanych z archiwizowaniem gwarowych danych mówionych w postaci korpusu językowego. Podstawą tego studium jest elektroniczny korpus języka mówionego mieszkańców Spisza (*Korpus Spiski*). Nie będziemy tu szczegółowo prezentować samej budowy korpusu ani też omawiać narzędzi informatycznych, gdyż na ten temat powstały już opracowania (por. Waldenfels, Woźniak 2016; Grochola-Szczepanek, Woźniak 2018b; Grochola-Szczepanek, Górski, Waldenfels, Woźniak 2019). Nasza uwaga zwrócona będzie na samo niestandardowe tworzywo korpusu, jakim jest wiejska polszczyzna mówiona. Pragniemy pokazać metodę gromadzenia danych gwarowych oraz ich zapisu, wskazując przy tym główne problemy i ograniczenia wynikające z nienormatywności danych mówionych oraz zastosowane rozwiązania.

Dane mówione w korpusach

Jak już wspomniano, elektroniczne korpusy językowe z możliwością przeszukiwania danych są obecnie najlepszą metodą archiwizowania różnych odmian języka, w tym także regionalnych i dialektalnych. Stanowią także bardzo użyteczne narzędzie badawcze współczesnego lingwisty.

Chociaż języki słowiańskie są dużym wyzwaniem dla lingwistyki komputerowej, to powstały już wielkie korpusy słowiańskie z wyszukiwarkami morfologicznymi, np. *Narodowy Korpus Języka Polskiego, Český národní korpus*[1]. Korpusy narodowe są referencyjne, powinny zatem uwzględniać także dane mówione (por. Pęzik 2012: 37), zwykle jednak są dużymi kolekcjami języka pisanego[2]. Dane mówione stanowią w nich niewielki procent, tworząc małe zbiory dołączone jako podkorpusy, np. podkorpus języka mówionego w *Narodowym Korpusie Języka Polskiego*, podkorpus gwarowego języka mówionego w *Czeskim Narodowym Korpusie* (*Český národní korpus*). Poza tym występują także samodzielne bazy, np. korpus mówionego języka

[1] Najliczniejsze i najbardziej zróżnicowane zbiory zawierają narodowe korpusy angielskie (m.in. *British National Corpus*) oraz amerykańskie (m.in. *American National Corpus*).

[2] Problem małych zasobów danych mówionych dostrzegany jest szerzej w językoznawstwie (por. Bańko, Kłosińska 1994; Przybylska 2009).

słoweńskiego (GOS: *Referenčni govorni korpus slovenskego jezika*). Przewaga korpusów pisanych nad mówionymi wynika z prostej przyczyny: gromadzenie danych mówionych jest znacznie trudniejsze niż danych pisanych. Tworzenie korpusu języka mówionego wiąże się z koniecznością rejestrowania wypowiedzi u określonych respondentów oraz transkrybowania nagrań. Rejestrowanie rozmów jest wymagające od strony metodologicznej, logistycznej oraz etycznej. Zapisywanie spontanicznej mowy, która często wykracza poza zakres polszczyzny ogólnej lub narusza normę językową, stwarza wiele problemów podczas anotacji. W rezultacie opracowanie danych mówionych w celu przystosowania ich do udostępnienia w formie „przeszukiwalnego" korpusu jest procesem wieloaspektowym, bardzo czasochłonnym i kosztownym.

Korpusy zawierające dane mówione to tzw. *speech corpora* lub *spoken corpora*. Pierwszy typ korpusów, czyli korpusy mowy, to zasoby nagrań wykorzystywane m.in. do badania zjawisk fonetycznych (Baza Mazak), do rozpoznawania mówcy w ramach lingwistyki kryminalistycznej (por. Klessa i in. 2013) lub automatycznego rozpoznawania mowy (por. Wagner i in. 2015). Korpusy języka mówionego (*spoken corpora*) to zasoby transkrypcji powstałych na podstawie nagrań, zwykle łączące warstwę tekstową z dźwiękową. Największym zbiorem polszczyzny mówionej jest podkorpus mówiony w NKJP, który udostępnia jednak tylko same transkrypcje, bez możliwości odsłuchiwania nagrań. Dane mówione stanowią dziesiątą część zrównoważonego korpusu, czyli 30 mln słów tekstowych, przy czym ponad 27 mln segmentów to dane mówione ze stenogramów posiedzeń Sejmu lub sejmowych komisji śledczych (czyli raczej dane czytane), 1,9 mln to dane mówione konwersacyjne, 900 tys. – dane mówione medialne, transkrypcje audycji radiowych i telewizyjnych (por. Pęzik 2012: 37–47). Wyniki w korpusie mówionym NKJP połączone są z następującymi metadanymi: tytuł, źródło, data nagrania, płeć rozmówcy, wiek i wykształcenie. Drugim co do wielkości zasobem danych mówionych jest korpus konwersacyjny *Spokes*, powstały w ramach polskiej infrastruktury CLARIN (por. Pęzik 2014). Podstawowy zbiór zawiera ponad 2 miliony słów tekstowych swobodnego dyskursu mówionego, który przedstawiony jest w formie transkrypcji oraz nagrań, podających takie same zmienne, jak NKJP. Jako trzecie opracowanie języka mówionego wymieńmy korpus mowy mieszkańców Spisza, do którego odnosi się niniejszy tekst. *Korpus Spiski* to „pierwszy elektroniczny korpus polskiej gwary, spełniający wszystkie wymagania, jakie współcześnie stawia się tego typu narzędziom" (Grochola-Szczepanek i in. 2019). Projekt powstał w Instytucie Języka Polskiego PAN w Krakowie w latach 2015–2019[3]. Korpus zawiera około 2 mln form tekstowych, opartych na nagraniach mowy, którą obecnie posługują się mieszkańcy regionu polskiego Spisza, bez względu na wiek oraz inne parametry. Wyszukiwarka pozwala na przeszukiwanie tekstów i nagrań według różnych kryteriów, m.in.: lematów, postaci tekstowej, określonych form gramatycznych, metadanych (identyfikator, wiek, płeć, miejscowość, wykształcenie, narodowość) oraz korespondującego fragmentu nagrania (por. Grochola-Szczepanek i in. 2019).

[3] Pełna nazwa projektu brzmi: *Język mieszkańców Spisza. Korpus tekstów i nagrań gwarowych*.

Rejestrowanie mowy mieszkańców wsi

Pozyskanie gwarowych danych mówionych wiąże się z koniecznością eksploracji na danym terenie. Badania polegają na nagrywaniu wywiadów z respondentami.

Podczas projektowania badań terenowych opracowuje się schemat zagadnień do rozmów, zawierający tematykę życia wiejskiego, m.in.: region i jego mieszkańcy, dzieciństwo, młodość, prace gospodarskie, dom i rodzina, zwyczaje i obrzędy, współczesność i zmiany na wsi. Każdy temat zawiera szereg szczegółowych pytań. Odpowiednie zestawy zagadnień dobierane są podczas wywiadu w zależności od wieku informatora oraz jego życiowych doświadczeń. Eksplorator po krótkim wywiadzie biograficznym z respondentem podejmuje decyzję o tematach rozmowy. Ze starszymi mieszkańcami rozmawia się zwykle o ich młodości, pracach gospodarskich, zwyczajach wiejskich i rodzinnych oraz zmianach, które zaszły na wsi. Z osobami młodszymi łatwiej rozmawia się o szkole, pracy poza wsią oraz o ich zainteresowaniach. Powtarzające się wywiady na te same tematy pozwalają zaobserwować zróżnicowanie języka informatorów oraz zweryfikować pozyskane informacje na temat poruszanych kwestii. Warto zaznaczyć, że schemat zagadnień odgrywa rolę tylko pomocniczą podczas wywiadów. Informatorzy często sami wychodzą poza poruszane zagadnienia, opowiadając ciekawie o swoich zainteresowaniach, przygodach. Głównym założeniem wywiadów nie jest ścisłe trzymanie się zagadnień szablonowych, lecz dążenie do tego, aby rozmówcy posługiwali się w miarę swobodnym językiem, którym mówią na co dzień.

Przed przystąpieniem do badań terenowych eksploratorzy muszą zapoznać się z procedurą prowadzenia i nagrywania wywiadów z respondentami, aby rezultaty pracy w terenie spełniały przyjęte wymogi, związane m.in. z reprezentatywnością danych, jakością i etycznością nagrywania. Eksploratorzy przechodzą szkolenie z zakresu zasad prowadzenia wywiadu i rejestrowania danych z użyciem dyktafonu. W ramach szkolenia odbywa się także tzw. badania obserwacyjne, podczas których doświadczony badacz rozmawia z respondentem, a potencjalni eksploratorzy biorą udział jako obserwatorzy. Bardzo pożyteczna jest stała wymiana doświadczeń eksploratorów po każdym pobycie w terenie. Eksploratorzy muszą być obeznani z gwarą oraz kulturą wiejską. Idealnymi kandydatami na eksploratorów są autochtoni czynnie posługujący się gwarą. Respondenci nie czują skrępowania przed takimi osobami i dzięki temu mogą swobodnie posługiwać się kodem gwarowym. W przeciwnym wypadku informatorzy mogą dążyć do eliminowania cech gwarowych i prób przechodzenia na kod osoby prowadzącej rozmowę. Wypowiedzi stają się wtedy nienaturalne i niespójne, co w rezultacie wpływa niekorzystnie na wynik badania. Z naszych obserwacji wynika, że kod eksploratora ma duży wpływ na kod respondenta. Ważnymi cechami eksploratorów są także otwartość, wchodzenie w interakcje z badanymi oraz zachowanie kultury osobistej i taktu wobec badanych.

Na wstępie rozmowy każdy potencjalny respondent informowany jest o celu badania oraz o tym, że wywiad jest rejestrowany. Informatorzy proszeni są także o pisemne wyrażenie zgody na wywiad i wykorzystanie nagrań z ich udziałem w korpusie. Każdemu rozmówcy pozostawia się specjalny list z informacją o projekcie, danymi kontaktowymi oraz podziękowaniami za udział w badaniu.

Obligatoryjnym elementem każdego nagrania jest wywiad socjologiczny, w którym odnotowuje się informacje o każdym respondencie. Do najważniejszych danych należą: rok urodzenia, płeć, miejsce zamieszkania, wykształcenie. W procesie gromadzenia metadanych zwraca się uwagę na kwestię prywatności informatorów. W celu zapewnienia poufności osobom uczestniczącym w nagraniach, dokonuje się anonimizacji: część metadanych oraz wrażliwe fragmenty nagrań nie są upubliczniane. Każdemu interlokutorowi przypisuje się tzw. identyfikator, pozbawiony imienia i nazwiska.

Podstawową metodą badań jest wywiad indywidualny z mieszkańcami regionu, trwający średnio ok. 40 min. Średnia liczba respondentów w poszczególnych wsiach waha się od 20 do 35. Uzależnione to jest od ilości oraz jakości pozyskanego materiału nagraniowego. Dobór rozmówców jest celowy. Wywiady muszą być prowadzone z respondentami w różnym wieku, ale za priorytetowe uważa się nagrania z osobami starszymi, u których gwara jest najlepiej zachowana. Oprócz wywiadów indywidualnych stosuje się tzw. diady, czyli rozmowy z dwoma osobami, głównie z małżeństwem. Wywiady w większym gronie rozmówców prowadzone są sporadycznie, głównie w grupach: rodzinnych (rodzice, dzieci, dziadkowie), mężczyzn lub kobiet w tym samym wieku, mieszanych pod względem wieku i płci. Rozmowy w grupie są bardziej ożywione, naturalne i spontaniczne. Trzeba mieć jednak świadomość, że anotacja danych mówionych z wywiadów grupowych jest trudniejsza. Nakładające się głosy rozmówców ostatecznie zaburzają czytelność fragmentów nagrania oraz przysparzają problemów podczas transkrypcji. Wyraźne nagranie jest bardzo ważne w korpusie danych mówionych. Z tego względu w badaniach korpusowych dominującą metodą pozyskiwania materiałów jest wywiad indywidualny.

Cały wywiad rejestrowany jest za pomocą dyktafonu cyfrowego oraz archiwizowany na komputerach, dysku zewnętrznym oraz serwerze. Pliki dźwiękowe zapisywane są w formacie WAV. Rozmiary tego typu plików są stosunkowo duże: 1 godzina nagrań to około 600 MB danych, ale format ten umożliwia bezstratny zapis dźwięku, co jest ważne zarówno ze względu na precyzję transkrypcji, jak i późniejsze użytkowanie w korpusie.

Z jakimi problemami można się zetknąć podczas badań terenowych? Zorganizowanie badań w terenie wymaga nawiązania kontaktów z lokalną społecznością, zwłaszcza z sołtysami, dyrektorami szkół, księżmi i regionalistami. W ten sposób łatwiej dotrzeć do potencjalnych respondentów. Z reguły nie da się przewidzieć, czy osoba, z którą chcielibyśmy przeprowadzić wywiad, skłonna będzie wziąć udział w badaniu oraz jak przebiegnie sama rozmowa. Informacja o nagrywaniu oraz konieczność podpisania zgody może wywołać negatywną decyzję u potencjalnego informatora. Eksplorator musi wykazać zrozumienie dla osób, które odmawiają udziału w wywiadzie, rezygnują z podpisania zgody na wywiad lub żądają wycofania wywiadu. Kolejnym problemem jest odczuwanie skrępowania przez informatora podczas rozmowy. Fakt nagrywania rozmowy jest stresujący dla wielu rozmówców. Niektórzy nie czują się pewnie, kiedy rozmawiają z kimś obcym, używającym kodu standardowego. Zauważa się wtedy, że nie mówią swoim naturalnym kodem, lecz próbują przechodzić na kod ogólny. Niekiedy można zaobserwować, że ten sam respondent inaczej rozmawia z domownikami, a inaczej z eksploratorami. Pomocny

w takiej sytuacji jest eksplorator autochton lub wsparcie osoby o zainteresowaniach regionalistycznych, pochodzącej z otoczenia respondenta. Wywiady z mieszkańcami wsi nagrywa się w ich domach. Podczas rozmowy mogą zdarzać się różnego typu zakłócenia, np. przerwanie wywiadu, pojawienie się nowych osób, nakładanie się głosów, hałas.

Reprezentatywność danych

Planowanie prac archiwizacyjnych języka mieszkańców danego regionu lub gwary wiąże się nie tylko z ustaleniem celu i sposobu badań. Ważnym zagadnieniem jest reprezentacja danych mówionych. Przyjmując założenie, że podstawą korpusu języka mówionego (podobnie, jak w korpusach tekstów pisanych) jest reprezentatywna próbka danej wspólnoty językowej, dane mówione muszą obejmować proporcje kobiet i mężczyzn, rozmówców starszych i młodszych, osób z wykształceniem i bez wykształcenia. W ten sposób otrzymujemy materiały zróżnicowane pod względem socjalno-demograficznym. Podejmowanie badań mowy mieszkańców wsi wiąże się także z dokumentacją autentycznej gwary, co w tradycyjnej dialektologii przekłada się na rejestrowanie wypowiedzi tylko starszych i niewykształconych respondentów oraz wykluczanie osób młodszych i wykształconych z badań. Rzeczywisty kod gwarowy najlepiej zachowany jest u najstarszych reprezentantów wiejskich. Młodsi mieszkańcy wsi używają kodu mieszanego, będącego połączeniem gwary i naleciałości z języka ogólnego. Powstaje zatem pytanie, czy w badaniach korpusowych możliwe są do zrealizowania obydwa wymienione cele: reprezentacja danych oraz dokumentacja najstarszej warstwy gwary. Próba połączenia dwóch celów, które wykluczają się w pewnym stopniu, nie jest łatwym zadaniem. Dobór respondentów musi uwzględniać dwa zasadnicze kryteria metodologiczne: ma zapewniać reprezentatywność próby, ale także dokumentować charakterystyczną dla regionu mowę. Aby spełnić obydwa warunki, badaniami należy objąć przedstawicieli różnych generacji oraz zadbać o wyraźną nadreprezentację osób najstarszych. W ten sposób próbka danych w korpusie pozwala na badania możliwie czystego systemu gwarowego u najstarszych mówców, jak i procesu zmian cech gwarowych pod wpływem języka ogólnego u przedstawicieli średniego i najmłodszego pokolenia. Warto zaznaczyć, że nawet wyrównany udział reprezentantów wszystkich pokoleń w badaniu nie daje gwarancji uzyskania danych mówionych równych pod względem ilościowym w poszczególnych grupach wiekowych. Długość wywiadu oraz liczba uzyskanych danych w wywiadzie silnie koreluje z wiekiem mówcy. Wywiady ze starszymi informatorami trwają zwykle dłużej niż z osobami młodszymi. Jest w nich znacznie mniej pytań eksploratora, a więcej wypowiedzi samego informatora, natomiast w rozmowie z młodszymi eksplorator musi częściej zadawać pytania, ponieważ informatorzy stosują krótkie, nierozbudowane odpowiedzi. Według statystyk, pytania eksploratora padają średnio 2 razy częściej w wywiadzie z młodym informatorem niż ze starszym (por. Grochola-Szczepanek, Woźniak 2018a: 81). Stan pełnego zrównoważenia w przekrojowym badaniu języka mieszkańców wsi jest więc trudny do osiągnięcia.

Jak zapisać niestandardowy kod językowy?

Tworzenie korpusu językowego na danych gwarowych wiąże się z koniecznością zapisu nagrań w postaci tekstu. Przystępując do tego zadania, należy zadać sobie pytanie: w jakim stopniu badana gwara odbiega od języka standardowego i jakiego rodzaju odmienności występują w zarejestrowanej mowie. Wiedza ta potrzebna jest do opracowania spójnego systemu anotacji, który przystosowany jest do kodu standardowego i musi zmierzyć się z odmiennym wariantem języka, jakim jest gwara.

Dokonanie transkrypcji nagrań, potrzebnych do stworzenia korpusu, jest zadaniem dosyć złożonym. Na specyfikę pracy wpływa zarówno fakt, że tekst ma posłużyć do zbudowania korpusu, jak i sama istota gwary – systemu językowego odrębnego w stosunku do języka ogólnego.

> System anotacji musi spełniać określone reguły – zarówno jeśli chodzi o kwestie językowe, jak i techniczne. Z punktu widzenia językowego powinien umożliwiać spójną reprezentację wszystkich pożądanych cech gwary, pozwalać na rozdzielenie wypowiedzi eksploratora i poszczególnych informatorów oraz na uwzględnianie wszelkich dodatkowych informacji istotnych dla transkrypcji, np. objaśnień wyrazów dyferencyjnych czy uwag dotyczących nietypowych zdarzeń podczas wywiadu. Od strony technicznej potrzeba, aby transkrypcja sporządzana była przy zastosowaniu jednoznacznych reguł i była możliwa do przetwarzania przez automatyczne narzędzia, służące do tworzenia elektronicznego korpusu (Grochola-Szczepanek, Woźniak 2018b: 278).

Istotne jest także, by format danych spełniał normy któregoś ze standardów stosowanych do przechowywania danych językowych. Zapewnia to stabilność danych, ważną z perspektywy upływu czasu i ułatwia późniejsze korekty oraz wykorzystywanie ich w innych projektach.

Projektując korpus gwarowy, można rozważać różne formy notacji, np. transkrypcję fonetyczną (w alfabecie slawistycznym lub IPA), zapis wykorzystujący znaki z języka standardowego, ale oddający gwarową wymowę lub zapis standaryzowany (tabela 1).

Tabela 1. Potencjalne warianty zapisu nagrań gwarowych

fonetyczny (IPA)	*a prɨndzʲij ɛɛ vɛsɛluntkɔ tacjɛ zrɔbjɨwɔ lɨm cjɛd zabjɨli ɛfiŋkɛ i ɔf͡tsɛ zabjɨli ji i zrɔbjɨwɔ ɛɛ mjɛlɔnɨf tɔ bɨwa jɛdna vjɛ͡tsɛʒo lɨm ŋɛ bɨwɔ tag daŋuv jak tɛrɔs tɛ daŋa t͡sɔ fila dajum gɔtuvanɛ t͡sɔ aŋi mawɔ t͡sɔ zɛ stɔwu ludʑ͡ɛ jɛdʑ͡ɔm ŋʲiɨ*
gwarowy	*a pryndzyj sie weselóntko takie zrobiyło lym kied zabiyli świnke i owce zabiyli i i zrobiyło sie mielónyf to była jedna wiecerzo lym nie było tak daniów jak teroz te dania co fila dajóm gotuwane co ani mało co ze stołu ludzie jedzóm niy?*
standaryzowany	*a prędzej się weselątko takie zrobiło lem kiedy zabili świnkę i owcę zabili i i zrobiło się mielonych to była jedna wieczerza lem nie było tak dań jak teraz te dania co chwila dają gotowane co ani mało co ze stołu ludzie jedzą nie?*

W przypadku korpusu językowego najbardziej optymalnym rozwiązaniem okazuje się standaryzowany zapis ortograficzny. Tylko zastosowanie znormalizowanej ortografii pozwala na użycie istniejących narzędzi do lematyzacji i anotacji morfosyntaktycznej (zaprojektowanych na potrzeby polszczyzny ogólnej). Chociaż

zapis standaryzowany odbiega od rzeczywistej wymowy, to bardzo ułatwia tworzenie, a następnie korzystanie z korpusu. Nie zmienia to faktu, że dostarczone równolegle z transkrypcją nagrania pozwalają użytkownikom korpusu zapoznać się z rzeczywistą wymową. Fonetyczny zapis okazuje się nieprzydatny w korpusie jeszcze z dwóch powodów. Po pierwsze, korpus jest narzędziem przeznaczonym dla szerokiego grona użytkowników, a transkrypcja fonetyczna czytelna jest tylko dla wąskiej grupy osób. Po wtóre, zapis nagrań gwarowych w formie tekstu jest zadaniem niezwykle czasochłonnym. Transkrypcja fonetyczna wydłuża pracę, ponadto jest bardziej wymagająca dla osób dokonujących zapisu.

Znormalizowany zapis gwarowych danych na potrzeby korpusu opiera się na zasadzie praw głosowych (por. Grochola-Szczepanek i in. 2019: 171–172). Jeżeli forma gwarowa i ogólna są kontynuantami tych samych morfemów, to wyraz przyjmuje zapis standardowy. Jeśli natomiast są kontynuantami odmiennych morfemów, to wyraz notowany jest jako forma standaryzowana. Wyrazy standaryzowane posiadają w korpusie dwie formy zapisu w osobnych warstwach: ogólną (standaryzowaną) oraz gwarową (zgodną z wymową gwarową). Dla przykładu formy *biydny*, *cekać*, *pón* zapisywane są w formie standardowej *biedny*, *czekać*, *pan*, gdyż w obydwu odmianach języka nazwy kontynuują ten sam morfem, a odmienny kształt fonologiczny gwarowych morfemów można wyjaśnić za pomocą praw głosowych. Natomiast formy *baluwacka*, *myślałaf*, *zogłówecek* sprowadzane są do wersji standaryzowanej *balowaczka*, *myślałam*, *zagłóweczek* oraz dodatkowo notowane zgodnie z wymową gwarową. Innymi słowy, jeśli odmienność form polega tylko na występowaniu prostych, regularnych zmian fonetycznych, jak np. mazurzenie, występowanie samogłosek pochylonych, realizacja samogłosek nosowych czy przejście wygłosowego *-ch* → *-f* (*na tyf staryf nogaf* 'na tych starych nogach'), notuje się je w wersji standardowej, bez pokazywania tych zmian w zapisie. Jeśli natomiast formy gwarowe reprezentują głębsze zmiany, np. morfologiczne (*myślałaf*) lub stanowią oryginalne nazwy, nienotowane w języku ogólnym (*baluwacka*, *zogłówecek*), zapisywane są podwójnie, czyli w wersji standaryzowanej oraz gwarowej, np. *balowaczka//baluwacka*, *myślałam//myślałaf*, *zagłóweczek//zogłowecek*.

Konwencje zapisu

Gwara posiada szereg cech odróżniających ją od języka standardowego. Wszystkie te cechy powinny być oddane w transkrypcji (a co za tym idzie – w gotowym korpusie). Aby to umożliwić, stosuje się określone konwencje zapisu oraz specjalne symbole. Wydziela się grupy wyrazów, różniące się stopniem zbliżenia do języka ogólnego, i traktuje się je odmiennie w anotacji: 1) formy identyczne z językiem ogólnym lub mające regularne zmiany fonetyczne – sprowadzane są do postaci ogólnej, np. *sopa* → *szopa*, 2) wyrazy morfologicznie odmienne – zapisywane w obu wersjach – standardowej i gwarowej – z oddzielającym je znakiem //, np. *babów* → *bab//babów*, 3) formy mające odpowiedniki w języku ogólnym, ale różniące się semantycznie – zapisywane w wersji ogólnej oraz sygnalizowane znakiem ^, np. *babka* 'kopa siana' → ^*babka*, 4) leksemy niewystępujące w języku ogólnym, znane jedynie z gwary – znakowane są symbolem #, sprowadzane do postaci standaryzowanej

i zapisywane w obydwu wersjach: standaryzowanej i gwarowej, np. *baluwacka* → *#balowaczka//baluwacka.*

Wyrazy z grupy 2 i 4 zapisuje się w korpusie w dwóch wersjach: ogólnej i gwarowej. Obydwa warianty zapisywane są przy użyciu znaków ortografii ogólnej. Poziom anotacji ogólnej tworzony jest sztucznie, natomiast warstwa gwarowa oddaje w przybliżeniu wszystkie cechy wymowy gwarowej i jest bliższa rzeczywistości. Zanotowanie w dwóch wersjach danych zmienionych morfologicznie oraz nieznanych w języku ogólnym daje możliwość porównania form standaryzowanych z ich brzmieniem gwarowym.

W mowie mieszkańców wsi występuje zjawisko łączenia końcówek osobowych czasowników z innymi częściami mowy, np. *jeszczem nie była*, *jużem był.* Z tzw. ruchomą końcówką, nazwaną także aglutynantem (NKJP), mamy do czynienia o wiele częściej niż w języku ogólnym. Zjawisko to przysparza trudności podczas automatycznej anotacji morfosyntaktycznej. Tagery nie radzą sobie z prawidłową identyfikacją wyrazów z dołączonym aglutynantem. Dodatkowym problemem w anotacji kodu gwarowego jest fakt, że aglutynanty występują jako elementy wolnostojące, np. *juz jef był*, *jesce zek nie była.* Z tego powodu podczas transkrypcji miejsca z aglutynantami muszą być specjalnie oznaczone (tabela 2).

Tabela 2. Anotacja aglutynantów

Aglutynant	W transkrypcji (postać standaryzowana//postać gwarowa)	W korpusie (warstwa ogólna//warstwa gwarowa)
połączony z częścią mowy	*już.em//juz.ef był,* *jeszcze.m//jesce.k nie była*	*jużem był//juzef był,* *jeszczem nie była//jescek nie była*
wolnostojący	*już em//jef był,* *jeszcze em//zek nie była*	*już em był//juz jef był,* *jeszcze em nie była//jesce zek nie była*

W transkrypcji kodu gwarowego można wyróżniać wiele innych zjawisk językowych, np. odmienną składnię, np. *poleciał ku moście*, jednostki wielowyrazowe typu *na despet* ‘na złość’, frazeologizmy, np. *ogrywanie majek* ‘zwyczaj związany z Zielonymi Świątkami’. W nagraniach zdarzają się ponadto nieprecyzyjne przywoływania z innych języków, np. *haroszo* lub miejsca z niepewną wymową, np. *wsićko : sićko* ‘wszystko’. Odpowiedni system znakowania w transkrypcji pozwala na odnotowanie charakterystycznych zjawisk oraz wszelkich wątpliwości.

W czasie odsłuchiwania nagrań oraz opracowywania transkrypcji pojawiają się wątpliwości dotyczące m.in.: odróżniania wyrazów *stricte* dyferencyjnych od zmienionych morfologicznie (np. *jakisi*, *poźryć*, *tyźnie*), skategoryzowania wyrazów ogólnych lub gwarowych (np. *gazdówka*, *odziewać*, *wleźć*), ortografii oryginalnych nazw własnych, nienotowanych w języku ogólnym, (np. *Sibicno Góra*, *Trybsianka*, *Winterłajt*), zapisu i charakterystyki gramatycznej wyrazów synsemantycznych (np. *nale*, *noji*). Wyrazy typu *jakisi*, *poźryć*, *tyźnie* ze względu na zmiany morfologiczne można uznać za wyrazy oryginalne, typowo gwarowe lub za zmienione pod względem morfologicznym formy ogólne: *jakiś*, *spojrzeć*, *tygodnie*. W procesie kwalifikacji wyrazów pomocne są słowniki ogólne. Nazwy *gazdówka*, *odziewać*, *wleźć*

ze względu na ograniczony zasięg lub przestarzały charakter można traktować jako formy gwarowe lub ogólne. Formy *nale*, *noji* charakterystyczne są dla języka mówionego, zwłaszcza opowiadania. Stanowią połączenia partykuły *no* i spójnika *ale* lub *i*. Pod względem brzmieniowym tworzą jedną jednostkę, dlatego zapisywane są łącznie. W oryginalnych nazwach własnych trudno niekiedy ustalić pochodzenie, a co za tym idzie – prawidłowy zapis, dlatego występują w zapisie zgodnym z brzmieniem.

Zasady transkrypcji mogą ulegać pewnym modyfikacjom w trakcie trwania procesu transkrybowania. Powiększająca się baza daje możliwość oglądania danych w większej reprezentacji oraz sukcesywnie coraz lepszą wiedzę o całym zbiorze. To z kolei prowadzi do pewnych zmian w sposobie zapisu i znakowania wariantywnych form. Przez cały czas muszą odbywać się spotkania zespołu, wspólne ustalenia, weryfikacje oraz szkolenia dla osób wykonujących transkrypcje w programie ELAN (por. Grochola-Szczepanek, Woźniak 2018b).

Warto nadmienić, że w zapisach robionych z odsłuchu zdarzają się miejsca błędnie zinterpretowane lub niepewne. W elektronicznym korpusie istnieje możliwość modyfikacji nawet w późniejszym czasie.

Główne etapy procesu anotacji tekstowej

Każde nagranie w procesie transkrypcji przechodzi szereg etapów. Do najważniejszych należą:

1) wybór nagrania według kryteriów: jakość nagrania, dobra gwara i ciekawy temat,
2) odsłuchiwanie i zapis przez anotatora zgodnie z przyjętymi zasadami,
3) pierwsza korekta wymienna pomiędzy anotatorami (sprawdzenie zapisów pod kątem podziału na segmenty, błędów literowych, poprawności zapisu i znakowania),
4) druga korekta (sprawdzenie poprawności zapisu, poprawności znakowania, dodanie objaśnień do wyrazów dyferencyjnych, uwag o nagraniu),
5) korekta techniczna (sprawdzanie wcześniej zrobionych partii zapisów i skorygowanie identyfikacji warstw lub podziału na segmenty),
6) korekta globalna (sprawdzanie w całej bazie wybranych form, realizacji i znakowania),
7) sprawdzanie po jakimś czasie miejsc wątpliwych (niektóre niejasne fragmenty nagrań udaje się zidentyfikować dopiero po długim okresie wielokrotnego przesłuchiwania przez wiele osób).

Anotacja tekstowa w praktyce

W wyniku transkrypcji powstają dwa poziomy zapisu: główny – standaryzowany oraz dodatkowy – gwarowy. Poziom pierwszy, nazwany w korpusie – ogólnym, powstaje przede wszystkim na potrzeby tagera i sprowadza wszelkie standardowe zmiany fonetyczne (np. mazurzenie, samogłoski pochylone, rozkład nosówek) do postaci ogólnej. Poziom gwarowy uwzględniony jest tylko dla oryginalnych typowo gwarowych nazw oraz form ze zmianami morfologicznymi, np. innym morfemem,

innym paradygmatem fleksyjnym, aglutynantem przyłączonym do innej części mowy lub wolnostojącym. Obydwa poziomy zapisywane są przy użyciu znaków ortografii polskiej. Poziom notacji ogólnej tworzony jest sztucznie, natomiast poziom notacji gwarowej jest bliższy rzeczywistym realizacjom, zachowuje wszystkie cechy wymowy gwarowej. Podczas transkrypcji obydwa poziomy zapisuje się w jednej warstwie, natomiast korpus pozwala na wybranie odpowiedniej warstwy. Poziom anotacji w warstwie gwarowej można w przyszłości uzupełnić także o zapisy wszystkich wyrazów, które obecnie mają tylko warstwę ogólną (formy wspólne z drobnymi zmianami fonetycznymi). Pozwoliłoby to śledzić zapis gwarowy segmentów w całości. Wszystkie etapy i warstwy zapisu można obejrzeć na przykładzie przywołanego wcześniej segmentu transkrypcji (tabela 3).

Tabela 3. Zapis w praktyce

1. Rzeczywista wypowiedź	**Zapis zgodny z wymową:** *a pryndzyj sie weselóntko takie zrobiyło lym kied zabiyli świnke i owce zabiyli i i zrobiyło sie mielónyf to była jedna wiecerzo lym nie było tak daniów jak teroz te dania co fila dajóm gotuwane co ani mało co ze stołu ludzie jedzóm niy?*
2. Transkrypcja w ELANIE	**Poziom ogólny i gwarowy łącznie:** *a prędzej się weselątko//weselóntko takie zrobiło lem//lym kiedy//kied zabili świnkę i owcę zabili i i zrobiło się mielonych to była jedna wieczerza lem//lym nie było tak dań//daniów jak teraz te dania co chwila dają gotowane//gotuwane co ani mało co ze stołu ludzie jedzą nie?*
3. Obecnie w korpusie (możliwość wyboru poziomu)	**Poziom ogólny:** *a prędzej się weselątko takie zrobiło lem kiedy zabili świnkę i owcę zabili i i zrobiło się mielonych to była jedna wieczerza lem nie było tak dań jak teraz te dania co chwila dają gotowane co ani mało co ze stołu ludzie jedzą nie?* **Poziom gwarowy (częściowy):** *a prędzej się weselóntko takie zrobiło lym kied zabili świnkę i owcę zabili i i zrobiło się mielonych to była jedna wieczerza lym nie było tak daniów jak teraz te dania co chwila dają gotuwane co ani mało co ze stołu ludzie jedzą nie?*
4. Potencjalne w korpusie (możliwość wyboru poziomu)	**Poziom ogólny:** *a prędzej się weselątko takie zrobiło lem kiedy zabili świnkę i owcę zabili i i zrobiło się mielonych to była jedna wieczerza lem nie było tak dań jak teraz te dania co chwila dają gotowane co ani mało co ze stołu ludzie jedzą nie?* **Poziom gwarowy (całościowy):** *a pryndzyj sie weselóntko takie zrobiyło lym kied zabiyli świnke i owce zabiyli i i zrobiyło sie mielónyf to była jedna wiecerzo lym nie było tak daniów jak teroz te dania co fila dajóm gotuwane co ani mało co ze stołu ludzie jedzóm niy?*

Podsumowanie

Archiwizowanie gwarowych danych mówionych w celu budowy korpusu językowego to proces złożony i wieloetapowy. Łączy z jednej strony gwarową, historyczną polszczyznę mówioną z nowoczesnym językoznawstwem korpusowym,

z drugiej – działania na różnych obszarach językoznawstwa i informatyki. Opiera się ponadto na zaangażowaniu i współpracy wielu osób: eksploratorów, respondentów, transkrybentów, lingwistów komputerowych, dialektologów, informatyków. Staraliśmy się tutaj omówić najważniejsze zagadnienia związane z dwoma etapami opracowania: rejestracją nagrań gwarowych w terenie oraz przetworzeniem warstwy brzmieniowej na tekstową.

Zaprojektowane badania terenowe, przebiegające według z góry ustalonych zasad, pozwalają pozyskać odpowiedni materiał językowy oraz wymagane dane o informatorach z zachowaniem zasad etycznych.

System anotacji kodu niestandardowego na potrzeby korpusu musi spełniać określone normy językowe oraz techniczne, m.in. zachowanie najważniejszych cech gwary i czytelny zapis. Wypracowane zasady są pewnego rodzaju kompromisem między właściwościami kodu niestandardowego a możliwościami narzędzi informatycznych oraz potrzebami korpusu. Trudności z notacją kodu niestandardowego wynikają z trzech głównych przyczyn:

1) braku wypracowanego spójnego systemu zapisywania gwar, które istnieją tylko w odmianie mówionej (normy zapisu języka ogólnego kształtowały się przecież przez wieki),
2) wariantywności współczesnej mowy mieszkańców wsi (kod mieszany: gwara + język ogólny jest powszechny zwłaszcza u młodszych),
3) konieczności użycia narzędzi informatycznych opracowanych dla języka standardowego (niezbędne byłyby narzędzia do konkretnego systemu gwarowego).

Rozwój gwarowych badań korpusowych jest dopiero w początkowym stadium. Należy mieć nadzieję, że metodologia wypracowana w *Korpusie Spiskim* przyczyni się do powstania innych gwarowych opracowań korpusowych. Archiwizacja danych mówionych w formie korpusu otwiera nowy rozdział badań nad językiem mieszkańców wsi, oparty nie na fragmentarycznie zapisywanych danych tekstowych, lecz na ciągłej żywej mowie.

Rozwiązanie skrótów

GOS – *Referenčni govorni korpus slovenskega jezika*, http://korpus-gos.net (dostęp: 07.02.2021).

NKJP – *Narodowy Korpus Języka Polskiego*, http://nkjp.pl (dostęp: 07.02.2021).

Bibliografia

Bańko M., Kłosińska A., 1994, *Polszczyzna mówiona nieobecna w słownikach*, [w:] *Współczesna polszczyzna mówiona w odmianie opracowanej (oficjalnej)*, red. Z. Kurzowa, W. Śliwiński, Kraków, s. 89–96.

Dunaj B., 1986, *Dialektologia a socjolingwistyka*, „Folia Linguistica" 12, s. 15–23.

Grochola-Szczepanek H., Górski R.L., von Waldenfels R., Woźniak M., 2019, *Korpus języka mówionego mieszkańców Spisza*, „LingVaria" LV/1, s. 165–180.

Grochola-Szczepanek H., Woźniak M., 2018a, *Badania korpusowe języka mieszkańców Spisza a czynnik pokoleniowy*, [w:] *Dialog pokoleń w języku i językoznawstwie*, red. E. Wierzbicka-Piotrowska, Warszawa, s. 79–90.

Grochola-Szczepanek H., Woźniak M., 2018b, *Transkrypcja języka mieszkańców wsi w aplikacji ELAN w Korpusie Spiskim*, [w:] *Historia języka, dialektologia i onomastyka w nowych kontekstach interpretacyjnych*, red. R. Przybylska, M. Rak, A. Kwaśnicka-Janowicz, Kraków, s. 267–278.

Klessa K., Wagner A., Oleśkowicz-Popiel M., Karpiński M., 2013, *Paralingua – A New Speech Corpus for the Studies of Paralinguistic Features*, „Procedia-Social and Behavioral Sciences" 95, s. 48–58.

Labocha J., 2012, *Pragmatyczne mechanizmy składni języka mówionego*, „Slavia Occidentalis" 69, s. 139–145.

Lewaszkiewicz T., 2017, *O zapisach fonetycznych polskiej i słowiańskiej mowy ludowej i potocznej*, „Gwary Dziś" 9, s. 183–197.

Przybylska R., 2009, *Badania nad polszczyzną mówioną a leksykografia*, [w:] *Polszczyzna mówiona ogólna i regionalna*, red. B. Dunaj, M. Rak, Kraków, s. 33–39.

Sierociuk J., 2009, *Zasoby fonograficzne Zakładu Dialektologii Polskiej Uniwersytetu im. Adama Mickiewicza i ich przydatność w badaniach procesów rozwojowych polszczyzny mówionej*, [w:] *Polszczyzna mówiona ogólna i regionalna*, red. B. Dunaj, M. Rak, Kraków, s. 179–188.

Wagner A., Bachan J., Klessa K., Demenko G., 2015, *Przegląd wybranych aspektów analizy prozodii mowy spontanicznej na potrzeby technologii mowy*, „Prace Filologiczne" LXVI, s. 271–298.

Waldenfels R. von, Woźniak M., 2016, *SpoCo – a simple and adaptable web interface for dialect corpora*, „Journal for Language Technology and Computational Linguistics" 31, s. 155–170.

Źródła internetowe

Baza Mazak, *Akustyczna baza danych gwar mazowieckich. Wokalizm*, http://www.bazamazak.uw.edu.pl/ (dostęp: 07.02.2021).

Český národní korpus, http://ucnk.ff.cuni.cz (dostęp: 07.02.2021).

GOS – *Referenčni govorni korpus slovenskega jezika*, http://korpus-gos.net (dostęp: 07.02.2021).

Korpus Spiski, *Język mieszkańców Spisza. Korpus tekstów i nagrań gwarowych*, http://spisz.ijp.pan.pl (dostęp: 07.02.2021).

NKJP – *Narodowy Korpus Języka Polskiego*, http://nkjp.pl (dostęp: 07.02.2021).

Pęzik P., 2012, *Język mówiony w NKJP*, [w:] *Narodowy Korpus Języka Polskiego*, red. A. Przepiórkowski, M. Bańko, R. Górski, B. Lewandowska-Tomaszczyk, Warszawa, s. 37–47, http://nkjp.pl/index.php?page=3&lang=0 (dostęp: 27.02.2021).

Pęzik P., 2014, *Spokes – a search and exploration service for conversational corpus data*, https://clarin-pl.eu/dspace/bitstream/handle/11321/47/spokes_pezik.pdf?sequence=5&isAllowed=y (dostęp: 10.01.2021).

Przepiórkowski A., Bańko M., Górski R.L., Lewandowska-Tomaszczyk B. (red.), 2012, *Narodowy Korpus Języka Polskiego*, Warszawa, http://nkjp.pl/index.php?page=3&lang=0 (dostęp: 27.02.2021).

Spokes-CLARIN, http://spokes.clarin-pl.eu/ (dostęp: 07.02.2021).

From recording to corpus, i.e. the method of archiving of the rural speech with using digital linguistics tools

Abstract

The article presents the method of archiving of the rural speech during the development of the electronic language corpus. Attention is focused on how to get spoken data and transcription of non-standard dialect code. It also presents the problems and limitations resulting from non-normative spoken data and the solutions applied. The recording and converting of spoken language data for corpus is a complex and multi-phase process. The data is obtained from recorded interviews with respondents. The developed system of spoken data transcription combines the properties of non-standard code, the capabilities of tools and needs of corpus.