LingVaria
Rok XIV (2019) | 1 (27) | s. 165–180
https://doi.org/10.12797/LV.14.2019.27.11

Helena Grochola-Szczepanek
Instytut Języka Polskiego Polskiej Akademii Nauk, Kraków
helena.grochola@ijp.pan.pl
https://orcid.org/0000-0002-1511-0486

Ruprecht von Waldenfels
Friedrich-Schiller-Universität, Jena
ruprecht.waldenfels@gmail.com
https://orcid.org/0000-0001-5822-5040

Rafał L. Górski
Instytut Języka Polskiego Polskiej Akademii Nauk, Kraków
rafal.gorski@ijp.pan.pl
https://orcid.org/0000-0003-4727-2639

Michał Woźniak
Instytut Języka Polskiego Polskiej Akademii Nauk, Kraków
michal.wozniak@ijp.pan.pl
https://orcid.org/0000-0001-9018-2204

KORPUS JĘZYKA MÓWIONEGO MIESZKAŃCÓW SPISZA[1]

Słowa klucze: korpus, język mówiony, dialektologia, gwara spiska
Keywords: corpus, spoken language, dialectology, Spisz dialect

1. Wstęp

W dokumentacji gwar stoimy zawsze przed dylematem, czy lepiej ocalić możliwie wiele z nich, godząc się na dość powierzchowny zapis, czy raczej dokonać bardzo dogłębnego zapisu wybranych miejsc. By rzec metaforycznie: czy lepiej zeskanować całą powierzchnię, czy raczej dokonać głębokich sondowań w wybranych miejscach. Dopowiedzmy, że ten dylemat nie dotyczy jedynie dialektologii lub językoznawstwa arealnego, ale szerzej – historii sztuki, botaniki, zoologii, geologii itp.

1 Publikacja została napisana wspólnie przez wymienionych autorów. Wkład współautorów jest równy i wynosi po 25%.

Nie ma łatwej odpowiedzi na powyższe pytanie, niemniej postaramy się dowieść, że bardzo szczegółowa dokumentacja gwary z niewielkiego terenu jest naukowo cenna, nawet gdyby miało to skutkować „białymi plamami" na innych obszarach.

W niniejszym artykule opisujemy cele i realizację projektu stworzenia dużego elektronicznego korpusu języka mówionego mieszkańców polskiego Spisza[2]. Jest to praca pionierska w polskiej dialektologii; od z górą stu lat wydaje się teksty gwarowe, od półwiecza dostępne są nagrania[3]. Opisywane przedsięwzięcie to jednak bardzo duży zbiór tekstów w dwu powiązanych postaciach, tj. nagrania i transkrypcji. Jest on wyposażony w wyszukiwarkę, która pozwala w ciągu sekund odnaleźć wszystkie wystąpienia szukanego słowa czy też formy fleksyjnej. Mamy już korpusy polszczyzny ogólnej (np. NKJP), historycznej (KorBa, por. Gruszczyński, Adamiec, Ogrodniczuk 2013; Derwojedowa i in. 2014). *Korpus Spiski* to pierwszy elektroniczny korpus polskiej gwary, korpus spełniający wszystkie wymagania, jakie współcześnie stawia się tego rodzaju narzędziom.

Korpus Spiski jest blisko powiązany z projektami o podobnym profilu:

– korpusem języka mieszkańców regionu Ustji (Rosja, obwód archangielski), [on-line:] http://parasolcorpus.org/Pushkino;
– korpusem języka rusińskiego, [on-line:] http://russinisch.de;
– korpusem dialektów z pogranicza Litwy, Białorusi i Rosji, [on-line:] http://www.trimcocorpus.de/spoco/.

Opracowania te łączą: podobne rozwiązania metodologiczne, przyjęte rozwiązania techniczne, a także wspólna (do pewnego stopnia) infrastruktura informatyczna.

2. Zasięg geograficzny

Badaniami korpusowymi objęto region Spisza w Polsce (15 wsi). Badania nie obejmują Spisza po stronie słowackiej, którego obszar jest dużo większy. Eksploracja na całym terenie na pewno byłaby potrzebna, ale zadanie to jest trudne do wykonania na tym etapie prac, z powodu ograniczeń finansowych i zespołowych.

---

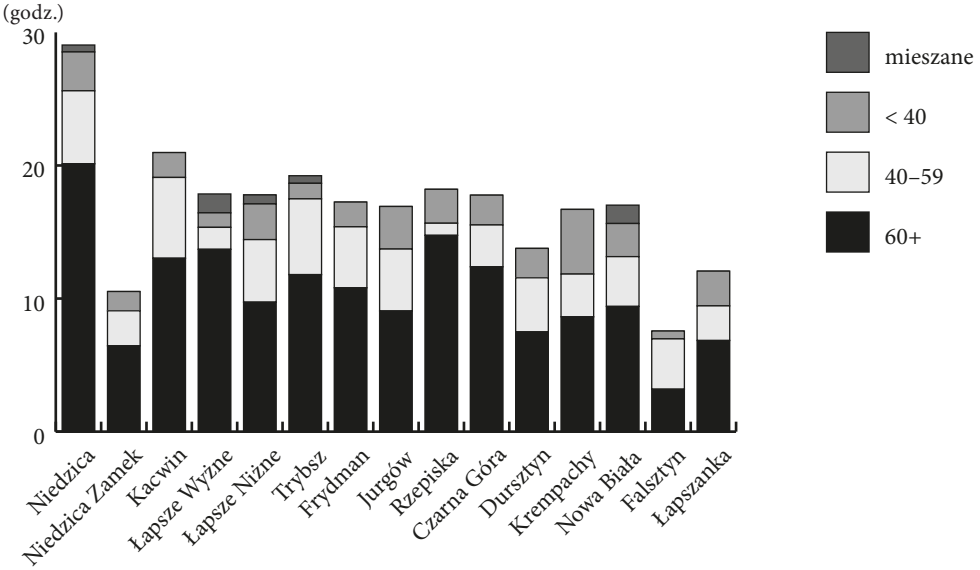
2 Projekt „Język mieszkańców Spisza. Korpus tekstów i nagrań gwarowych" jest finansowany przez NPRH w latach 2015–2019 (1bH 15 0166 83).

3 Istnieją już pewne zdygitalizowane zasoby gwarowe, jak np. *Akustyczna baza danych gwar mazowieckich* (Garczyńska 2013–2017) czy *Dialekty i gwary polskie. Kompendium internetowe* (Karaś 2010). Ten pierwszy jest jednak nastawiony wyłącznie na dane fonetyczne, drugi, choć łączy tekst z nagraniem, ze względu na bardzo ograniczoną wielkość nie może służyć do badań, lecz jedynie do popularyzacji wiedzy o dialektach. W literaturze omówiono prace nad korpusem wsi Maćkowce, ale do tej pory nie jest on publicznie dostępny (Krawczyk-Wieczorek 2012). Planowane są także prace nad *Korpusem Gwar Polskich* (Karaś, Kresa, Krawczyk-Wieczorek 2012).

3. Skład korpusu

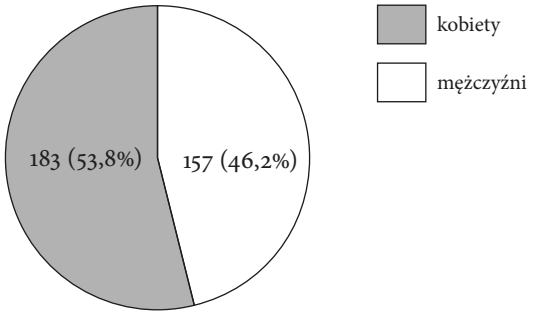
Korpus Spiski składa się z powiązanych ze sobą plików dźwiękowych i transkrypcji. Każdemu plikowi z nagranym wywiadem odpowiada jeden plik XML zawierający transkrypcję. Ogółem materiał liczy ponad 320 wywiadów przeprowadzonych z 340 informatorami, dających razem ok. 250 godzin nagrań[4].

Wykres 1. Długość nagrań w zależności od wsi i grupy wiekowej



* Grupa „mieszana" to kategoria przypisywana wywiadom, w których biorą udział osoby z różnych grup wiekowych.

Wykres 2. Udział kobiet i mężczyzn w nagraniach



4 Dysproporcja w liczbie wywiadów i informatorów wynika z tego, że niektóre wywiady przeprowadzono z dwojgiem lub kilkorgiem respondentów – zaznaczone są jako grupy mieszane (wykres 1).

W warstwie tekstowej korpus zawiera niemal 2 mln wyrazów, a ściślej mówiąc – tzw. segmentów (Przepiórkowski 2004)[5]. Segmentem może być wyraz lub znak interpunkcyjny, ponadto niektóre wyrazy są rozdzielane na trzy elementy, np. forma czasownika *chciałbyś* reprezentowana jest osobno przez segmenty: *chciał*, *by* i *ś*, zlematyzowane jako *chcieć*, *by* i *być*. Ten ostatni to – według terminologii NKJP – aglutynant: ruchomy morfem mogący dołączać się do innych części mowy poza czasownikiem. Wyniki wyszukiwania w korpusie wyświetlają się jako elementy dłuższych odcinków nagrania, które nazwane są segmentami transkrypcji. Odcinki te są wydzielane w trakcie transkrypcji, są fragmentem wypowiedzi respondenta. Segment w tym znaczeniu to w przybliżeniu odpowiednik zdania. Korpus składa się z blisko 90 tys. takich segmentów.

3.1. Reprezentatywność i zrównoważenie

Korpus językowy ma być reprezentatywną próbką mowy pewnej wspólnoty językowej. Zostawmy na boku (skądinąd bardziej skomplikowane) zagadnienie reprezentatywności korpusu tekstów pisanych i skupmy się na danych mówionych (Górski, Łaziński 2012). W tym bowiem przypadku reprezentatywność można rozumieć tak samo, jak próbkowanie w celu badania opinii publicznej, a więc, mówiąc w uproszczeniu, oczekiwalibyśmy, by próbka zawierała badaną społeczność w miniaturze, tj. by proporcje kobiet i mężczyzn (por. wykres 2), ludzi młodych i w wieku podeszłym, z wykształceniem wyższym i bez niego – były takie same, jak w całej społeczności. Tylko takie podejście pozwala na stworzenie „uśrednionego" obrazu języka. Świadomie jednak dokonaliśmy odstępstwa od takiego modelu. Mianowicie w doborze informatorów dążyliśmy do nadreprezentacji osób starszych, a więc takich, wśród których gwara jest zachowana najlepiej i ma najmniej cech przejętych z języka ogólnopolskiego. Duża liczebność tej grupy zapewnia względną obfitość danych, które dają dostęp do możliwie czystego systemu gwarowego. Z drugiej strony nagrania przedstawicieli średniego i najmłodszego pokolenia pozwalają obserwować procesy utraty cech gwarowych pod wpływem języka ogólnego. Co więcej, istotne jest uchwycenie gwary takiej, jaką ona naprawdę jest – również w komunikacji młodszych pokoleń.

Tak więc istotą prowadzonych badań korpusowych jest zarchiwizowanie języka, którym obecnie posługują się mieszkańcy wsi spiskich bez względu na wiek, wykształcenie oraz inne parametry. Podejście to różni się od tradycyjnych badań dialektologicznych, nastawionych z reguły na rejestrowanie tylko najstarszej warstwy gwary i wykluczanie młodszych lub wykształconych respondentów z badań (AJPP II:

5 Zwracamy uwagę, że w artykule pojawia się podobny termin: *segment transkrypcji* (zdefiniowany poniżej) mający inne znaczenie.

18). Z punktu widzenia rejestrowania gwary spiskiej istotne jest to, aby respondenci swobodnie posługiwali się podczas wywiadu językiem, którym mówią na co dzień. Mamy tutaj na myśli gwarę lub gwarę z niewielkimi naleciałościami z języka ogólnego. Dobór respondentów był zatem podyktowany dwoma zasadniczymi względami metodologicznymi: miał zapewnić reprezentatywność próby, ale także dokumentować charakterystyczną dla regionu mowę, czyli gwarę spiską. Mamy świadomość, że te dwa cele się w pewnym stopniu wykluczają. Eksploracja uwzględniająca wyżej wymienione kryteria jest zatem z założenia nie w pełni precyzyjna, projekt stanowi jednak kompromis między dwoma obrazami współczesnej gwary spiskiej i umożliwia badanie ich obu. Pomimo więc udziału w badaniach reprezentantów wszystkich pokoleń należy wyraźnie zaznaczyć, że dane w poszczególnych grupach wiekowych nie są równe pod względem ilościowym. Stan zrównoważenia w przekrojowym badaniu języka mieszkańców wsi jest trudny do osiągnięcia, właściwie nieosiągalny. Można jedynie mówić o próbie dążenia do zrównoważenia danych (por. tabela 1).

Tabela 1. Liczba segmentów w zależności od grupy wiekowej

Grupa wiekowa	Poniżej 40 lat	40–59 lat	Powyżej 59 lat
Liczba segmentów	281 632 (14,4%)	510 337 (26%)	1 168 900 (59,6%)

Warto podkreślić, że opisywany korpus dokumentuje stan z lat 2015–2018. Nie zawiera on żadnych tekstów wcześniejszych, nawet jeśli istnieją takie nagrania, zapewne też nie zostaną już powtórzone tak szeroko zakrojone prace dokumentacyjne. Jest to więc korpus ściśle synchroniczny, niedający wglądu w przeszłość gwary, ale dokumentujący pewien punkt w jej dziejach.

4. Etapy pracy

Opracowanie korpusu niestandardowego języka mówionego rozłożone jest na kilka etapów: gromadzenie materiałów (nagrywanie i archiwizowanie rozmów z respondentami), przetwarzanie danych (transkrypcja, znakowanie morfosyntaktyczne), wreszcie zapisanie ich w postaci bazy danych, która może być odpytywana przez wyszukiwarkę korpusową.

4.1. Badania terenowe

Materiały gwarowe zostały zgromadzone przez zespół eksploratorów w czasie badań terenowych we wszystkich 15 miejscowościach spiskich w Polsce. Dodajmy, że to zadanie właściwie nie różni się od tradycyjnych zadań dialektologów eksploratorów.

Rozmowy były nagrywane jawnie. Oczywistą konsekwencją takiej procedury jest fakt, że rezygnuje się ze spontanicznych dialogów pomiędzy użytkownikami gwary; przeciwnie – nagrywane teksty to wypowiedzi raczej o narracyjnym charakterze, w większości wypadków pozbawione interakcji rozmówców.

Respondenci zostali poproszeni o pisemne wyrażenie zgody na udział w badaniu oraz na wykorzystanie w korpusie nagrań z ich udziałem. Materiały dźwiękowe zarejestrowaliśmy przy pomocy dyktafonów Olympus LS-12 i Olympus LS-14, w formacie WAV[6].

Podczas badań terenowych w latach 2015–2018 zarejestrowaliśmy wywiady z ponad 600 mieszkańcami Spisza, co przekłada się na ok. 400 godzin nagrań, z czego do transkrypcji przeznaczyliśmy 250 godzin. Najwięcej rozmów przeprowadziliśmy z osobami urodzonymi w latach 40. XX w. (137 respondentów). Najstarszą rozmówczynią była kobieta urodzona w 1915 r. we wsi Frydman, a najmłodszym rozmówcą – uczeń z Nowej Białej, urodzony w 2008 r. Średni wiek respondenta wynosi 58 lat (mediana – 61, odchylenia standardowe – 22,3).

Podczas gromadzenia danych wykorzystano metodę badań socjolingwistycznych, która uwzględnia wpływ czynników społecznych na sferę języka (Lubaś 1979; Dunaj 1986). Takie podejście do badań wynika z obserwowanego silnego zróżnicowania języka mieszkańców wsi (Grochola-Szczepanek 2013; Wyderka 2014). Na niejednolitość kodu respondentów wiejskich wpływa wiele czynników, np. wiek, płeć, wykształcenie, wykonywany zawód, pochodzenie, dłuższe pobyty poza wsią. Z każdym respondentem został zatem przeprowadzony wywiad socjologiczny, podczas którego odnotowano wszystkie dane, które mogą mieć istotny wpływ na kod językowy badanego. Metoda badań terenowych została omówiona szczegółowo w osobnym artykule (Grochola-Szczepanek 2017).

4.2. Transkrypcja

Anotacja kodu niestandardowego musi zmierzyć się z odmiennym systemem językowym, jakim jest gwara, występowaniem leksemów nieznanych w języku ogólnym oraz form wariantywnych, odmiennych fonetycznie lub morfologicznie. Na specyfikę transkrypcji wpływa również fakt, że teksty mają posłużyć do zbudowania korpusu. System anotacji nagrań gwarowych na potrzeby korpusu powinien zatem spełniać określone normy językowe oraz techniczne. Proces opracowania transkrypcji przedstawiliśmy szczegółowo w innym artykule (Grochola-Szczepanek, Woźniak 2018). Niniejszy rozdział omawia najistotniejsze kwestie dotyczące koncepcji i przebiegu tego procesu.

6 Świadomie rezygnujemy z formatu MP3, który pozwala zachować dane w znacznie mniejszym pliku, ponieważ nie nadaje się on do badań fonetycznych.

Zasadnicze pytanie, jakie musieli sobie zadać twórcy korpusu na etapie formułowania jego koncepcji, dotyczyło standardu transkrypcji. Nasuwały się trzy rozwiązania:

1) transkrypcja fonetyczna (już to w wariancie slawistycznym, już to IPA);
2) transkrypcja półfonetyczna;
3) zapis w znormalizowanej ortografii polszczyzny standardowej.

Ewentualnym czwartym rozwiązaniem byłoby opracowanie niezależnej ortografii spiskiej, jednak nie mogło być ono brane pod uwagę przy tworzeniu korpusu.

Rozwiązanie pierwsze, jakkolwiek kuszące, wiązało się z trudnościami. Przede wszystkim byłoby ono dużo bardziej wymagające dla osób dokonujących transkrypcji, gdyż musiałyby za każdym razem podejmować decyzję co do tego, jak w danej wypowiedzi głoska została zrealizowana. Dodajmy – decyzję bardzo często dyskusyjną. Równocześnie, skoro równolegle z transkrypcją dostarczamy nagranie, osoby zainteresowane fonetyką czy nawet laicy, nieznający alfabetu fonetycznego, mogą zapoznać się z rzeczywistą wymową informatora.

Z kolei znormalizowany zapis ortograficzny jest dość odległy od rzeczywistej wymowy, a nawet od jej pewnej idealizacji. I tak np. spółgłoska szczelinowa dziąsłowa bywa oddawana ortograficznie przez <s> lub <sz>, zależnie od tego, jak jest realizowana w odpowiednim słowie w kodzie ogólnym, choć w wymowie informatora jest to za każdym razem ta sama głoska.

Dwa czynniki zadecydowały jednak o przyjęciu takiego zapisu. Po pierwsze, ułatwia on przeszukiwanie korpusu, tam zaś, gdzie warstwa dźwiękowa ma znaczenie (fonetyka, fonologia, morfologia, ewentualnie ich połączenie z socjolingwistyką), badacz sięgnie do pliku dźwiękowego, powiązanego ze znormalizowanym zapisem ortograficznym. Po drugie, zastosowanie znormalizowanej ortografii pozwoliło wykorzystać narzędzia służące do anotacji morfologicznej polszczyzny ogólnej (por. rozdz. 4.3).

Zapis półfonetyczny nie został zastosowany, gdyż łączy wady obu rozwiązań, pozbawiony natomiast jest ich zalet. Z jednej strony bowiem nie pozwala wykorzystać istniejących narzędzi do analizy języka ani też nie ułatwia przeszukiwania, ponieważ oddaje niekonsekwencję wymowy, z drugiej strony – jak już sama jego nazwa wskazuje – stanowi tylko przybliżenie rzeczywistej fonetyki, w rzeczywistości oddaje ją bardzo słabo.

Jeśli chodzi o morfologię, zasady są nieco inne. Otóż morfemy, których formę odmienną od tej, jaka jest w kodzie standardowym, da się wywieść jedynie za pomocą praw głosowych, zapisujemy w ortografii standardowej. Innymi słowy, jeżeli morfem kodu spiskiego i kodu ogólnego jest kontynuantem tego samego morfemu, to przyjmuje on zapis standaryzowany. Jeśli natomiast są one kontynuantami odmiennych morfemów, to przyjmujemy zapis oddający formę morfemu spiskiego. I tak np. 1 os. lp. czasu przeszłego od czasownika *chodzić* będzie zapisana jako *chodziyłef*, nie zaś jako *chodziłem*, gdyż pierwszy morfem jest kontynuantem *-ech*, drugi zaś

kontynuantem *-em*. Natomiast przymiotnik *biała* będzie zapisany w formie standaryzowanej, gdyż zarówno postać kodu ogólnego, jak i kodu spiskiego kontynuują ten sam morfem; odmienny kształt fonologiczny morfemu da się wyjaśnić za pomocą prawa głosowego.

Tak więc mamy do czynienia z czterema różnymi przypadkami, z których każdy reprezentuje różny stopień zbliżenia do języka ogólnego i w związku z tym jest traktowany nieco odmiennie:

1) Formy identyczne z językiem ogólnym lub mające regularne zmiany fonetyczne (np. mazurzenie, samogłoski pochylone): w zapisie są one sprowadzane do postaci ogólnej.
2) Wyrazy morfologicznie odmienne – jednostki mające bezpośrednie odpowiedniki w języku ogólnym, ale różniące się innym morfemem bądź paradygmatem fleksyjnym: w transkrypcji są zapisywane w obu wersjach – standardowej i gwarowej – z oddzielającym je znakiem //.
3) Formy mające odpowiedniki w języku ogólnym, ale różniące się semantycznie: w transkrypcji są zapisywane w wersji ogólnej oraz sygnalizowane znakiem ^.
4) Leksemy niewystępujące w języku ogólnym, znane jedynie w gwarze: w transkrypcji są znakowane symbolem # i zapisywane zgodnie z brzmieniem. Na dalszym etapie prac są sprowadzane do postaci standaryzowanej.

Tabela 2. Przykłady zapisu i znakowania klas wyrazów w transkrypcji oraz ich liczba w korpusie

Klasa wyrazu	Przykłady (zgodnie z brzmieniem)	Zapis w transkrypcji	Symbol	Liczba segmentów
1	*bedzie*, *cysty*, *mlyko*	*będzie*, *czysty*, *mleko*	brak	1 844 353
2	*dałak*, *krzikła*, *robotów*	*dałam//dałak*, *krzyknęła//krzikła*, *robót//robotów*	//	116 516
3	*boisko*, *bywać*, *ślafrok*	*boisko*, *bywać*, *szlafrok*	^	35 086
4	*janglusek*, *lym*, *odziywacka*	*janglusek*, *lym*, *odziywacka* (w ostatecznej wersji: *angluszek//janglusek*, *lem//lym*, *odziewaczka//odziywacka*)	#	70 812

Wyrazy z grupy 2 i 4 są notowane w dwóch wersjach – ogólnej i gwarowej. Obydwie wersje są zapisywane przy użyciu znaków ortografii ogólnej. Poziom anotacji ogólnej jest tworzony sztucznie, natomiast poziom notacji gwarowej oddaje w przybliżeniu wszystkie cechy wymowy gwary spiskiej i jest bliższy rzeczywistości. Notowanie danych w dwóch wersjach pozwoli na oglądanie form w wersji gwarowej lub standaryzowanej. W transkrypcji znakowane są także inne odmienności gwarowe, m.in. jednostki wielowyrazowe (*młodzi panowie* 'państwo młodzi, nowożeńcy'), odmienna składnia (*ku moście*), wtrącenia z języków obcych <pridi> 'przyjdź' (z jęz. słowackiego), <dawaj sało>, sało 'słonina' (z jęz. rosyjskiego).

Odsłuchiwanie i zapisywanie nagrań to najbardziej żmudny proces w tworzeniu całego korpusu. Przepisanie 1 godziny nagrania to nakład ok. 40 godzin pracy osoby dokonującej transkrypcji. Transkrypcje były kilkakrotnie weryfikowane z nagraniem przez kilka osób. Warto podkreślić, że korekty nie kończą się wraz z zakończeniem transkrypcji. Zapisy robione z odsłuchu są narażone na liczne błędy, dlatego istnieje możliwość ich modyfikacji nawet w późniejszym czasie.

4.3. Znakowanie morfosyntaktyczne

Gotowe transkrypcje przechowywane są jako pliki XML, przy czym każdy wywiad zapisywany jest w osobnym pliku. Na tym etapie kończy się ich ręczna edycja, pracę zaś przejmują narzędzia automatyczne. Pierwszym z automatycznych etapów jest znakowanie morfosyntaktyczne, tj. przypisanie każdemu segmentowi postaci hasłowej (lematu) i znacznika charakterystyki gramatycznej. Za standard przyjęliśmy tu tagset (zbiór kategorii gramatycznych i odpowiadających im symboli, a także sposób ich zapisu) NKJP. Przykładowo, segment *akordeonie* będzie miał znacznik *subst:sg:loc:m3*, gdzie poszczególne kategorie, rozdzielone dwukropkiem, oznaczają odpowiednio część mowy, liczbę, przypadek i rodzaj.

Automatyczne znakowanie przeprowadzane jest dwuetapowo: wyrazy funkcjonujące w polszczyźnie standardowej anotowane są za pomocą tagera Pantera. Do takich zaliczamy również podzbiór wyrazów dyferencyjnych, które są homonimiczne z wyrazami standardowymi i są takimi samymi częściami mowy (trzecia klasa wyrazów według klasyfikacji omówionej w rozdz. 4.2). Pozostałe wyrazy dyferencyjne, z którymi dostosowany do języka ogólnego tager sobie nie radzi, znakowane są na podstawie dodatkowej bazy danych, utworzonej w ramach projektu w programie Kuźnia.

4.4. Tworzenie korpusu

Kolejnym krokiem jest przetworzenie plików dźwiękowych i tekstowych do postaci przeszukiwalnej przez program konkordancyjny. Ten etap również jest w pełni automatyczny i składa się z kilku stadiów:

1) podział plików dźwiękowych na odcinki odpowiadające segmentom transkrypcji;
2) przetworzenie danych tekstowych do pliku w formacie wymaganym przez system CWB[7];
3) uzupełnienie pliku o metadane;
4) anonimizacja danych osobowych;
5) dołączenie objaśnień do wyrazów dyferencyjnych;
6) zapisanie danych w postaci bazy danych CWB;
7) połączenie bazy danych z interfejsem sieciowym.

5. Standardowe narzędzia

Prace nad korpusem języka mówionego są wieloetapowe, a każdy z tych etapów wymaga dużego nakładu pracy i staranności, by zachować spójność danych. Wykorzystanie istniejących już narzędzi pozwala w dużej mierze przyspieszyć i uprościć cały proces. Przy budowie *Korpusu Spiskiego* korzystaliśmy z poniższych narzędzi:

1. ELAN – jest to program do anotacji zasobów multimedialnych, stworzony w Instytucie Psycholingwistyki Maxa Plancka w Nijmegen (Brugman, Russel 2004). Program ten jest szeroko wykorzystywany przy projektach związanych z archiwizacją i opracowywaniem danych mówionych. Stanowi podstawowe środowisko pracy anotatorów zajmujących się transkrypcją nagrań w projekcie. Umożliwia segmentację i wielopoziomową anotację wywiadów. ELAN zapisuje transkrypcje nagrań w formacie XML, najpopularniejszym i najszerzej wykorzystywanym standardzie zapisu danych językowych, natomiast nagrania przechowywane są w formacie WAV, również jednym z popularnych standardów zapisu dźwięku.
2. Tager Pantera. Decyzja, aby normalizować wypowiedzi informatorów do języka standardowego, była w dużej mierze podyktowana potrzebą wykorzystania narzędzia do anotacji morfosyntaktycznej. Umożliwia ono wyszukanie wszystkich wystąpień danego wyrazu, niezależnie od ich formy, bądź wyszukanie wszystkich wyrazów o danej charakterystyce fleksyjnej. Istniejące dla języka polskiego tagery dostosowane są wyłącznie do polszczyzny ogólnej, zatem niemożliwe byłoby ich użycie, jeśli transkrypcja nie zawierałaby form wyrazowych języka ogólnego. W projekcie spiskim wykorzystaliśmy tager Pantera, cechujący się najwyższą precyzją anotacji.

7 The IMS Open Corpus Workbench, [on-line:] http://cwb.sourceforge.net/.

3. Kuźnia – stworzone w Instytucie Podstaw Informatyki PAN narzędzie do tworzenia słowników fleksyjnych języka polskiego. Ponieważ tagery nie są w stanie poradzić sobie z dyferencyjną częścią leksyki gwarowej, anotacja morfosyntaktyczna słownictwa dyferencyjnego została przeprowadzona ręcznie przy pomocy Kuźni. Umożliwia ona przypisanie leksemom nieistniejącym w polszczyźnie ogólnej paradygmatów fleksyjnych, form hasłowych, a także objaśnień (użytych do budowy słowniczka wyrazów dyferencyjnych).

Wykorzystanie istniejących narzędzi ma wiele oczywistych zalet: przyspiesza pracę i ułatwia kontrolę poprawności danych, pozwala na przyjęcie standardowych rozwiązań (np. format XML, tagset NKJP), co zwiększa spójność danych, ułatwia ich wykorzystanie w innych projektach, a także zapewnia ich większą stabilność przy długotrwałym przechowywaniu (Waldenfels, Woźniak 2016). Jednocześnie należy zauważyć, że użycie standardowych narzędzi do niestandardowej odmiany języka często nie jest możliwe bez odpowiednich modyfikacji. Modyfikacje te były dwojakiego rodzaju:

a) dotyczące danych – jak wspomniano wyżej, aby możliwe było wykorzystanie tagera, musieliśmy na etapie transkrypcji przeprowadzić normalizację ortograficzną;
b) dotyczące narzędzi – program Kuźnia dostosowany jest do systemu fleksyjnego języka standardowego. Ponieważ gwara w pewnych miejscach od tego systemu odbiega, niezbędne były modyfikacje programu, które umożliwiły przypisanie paradygmatów gwarowych leksemom dyferencyjnym. W związku z tym, że w korpusie wyrazy te przechowywane są w dwóch wersjach: znormalizowanej i gwarowej, konieczne było też dodanie do Kuźni możliwości reprezentacji tego samego leksemu w dwóch odmianach. Zmiany te możliwe były dzięki temu, że Kuźnia to oprogramowanie otwarte, jej kod źródłowy jest udostępniany na licencji BSD.

6. Efekt

Podstawowym rezultatem projektu jest korpus tekstów mówionych. Jest on dostępny w Internecie pod adresem https://spisz.ijp.pan.pl. Na jego potrzeby został stworzony interfejs sieciowy, który jest zmodyfikowaną wersją projektu SPOCO (ibid.). Dostęp do korpusu obecnie zabezpiecza hasło, ostateczna wersja korpusu będzie udostępniana bez ograniczeń.

Korpus umożliwia wyszukiwanie segmentów transkrypcji – składających się z odpowiedniego wycinka nagrania i powiązanej z nim transkrypcji. Jest on oparty na standardzie CWB – szeroko stosowanym zestawie narzędzi do tworzenia i przeszukiwania korpusów, pozwalającym na tworzenie złożonych zapytań i wyposażonym w wiele funkcji, umożliwiających analizy ilościowe i jakościowe. Standard ten wykorzystuje język zapytań CQL, oferujący duże możliwości, ale jednocześnie

wymagający od użytkownika znajomości jego składni. Aby nie zmuszać użytkownika do żmudnej nauki, interfejs korpusu umożliwia budowanie zapytań z prostych bloków składowych – wystarczy wpisać poszukiwany wyraz w jedno z pól wyszukiwania. Interfejs oferuje cztery typy pól wyszukiwania: *forma* – umożliwia szukanie za pomocą postaci tekstowej wyrazu w wersji znormalizowanej, *leksem* – pozwala na wyszukiwanie wszystkich form o określonej postaci hasłowej, *forma gwarowa* – wyszukuje postać tekstową wyrazu w wersji gwarowej i *tag gramatyczny* – zapewnia możliwość wyszukiwania wyrazów za pomocą ich właściwości gramatycznych. Wyniki wyszukiwania można ograniczać za pomocą filtrów (dostępne są: *płeć*, *narodowość*, *wykształcenie*, *miejsce zamieszkania*, *rok urodzenia*, *informator*).

Opisany wyżej moduł „wyszukiwania podstawowego" pozwala na odpytywanie korpusu za pomocą podstawowych zapytań. Zaawansowane zapytania (np. wyszukiwanie wszystkich wyrazów, które różnią się formą standardową i gwarową) dostępne są dzięki „wyszukiwaniu zaawansowanemu", dającemu znacznie większe możliwości, ale wymagającemu od użytkownika znajomości składni CQL.

Rezultaty wyszukiwania prezentowane są w postaci listy segmentów transkrypcji, zawierających sekwencje zdefiniowane w zapytaniu. Każdy segment można odsłuchać i zapoznać się z transkrypcją zarówno w wersji znormalizowanej, jak i gwarowej. Strona wyników umożliwia prezentację wyników na dwa sposoby: widok podstawowy wyświetla całość segmentów, widok KWIC (Key Word in Context – słowo kluczowe w kontekście) dzieli segment na trzy kolumny: kontekst lewostronny, dopasowanie i kontekst prawostronny. Oba sposoby prezentacji umożliwiają też sortowanie wyników, wyświetlanie skojarzonych z segmentem metadanych i wyświetlanie szerszego kontekstu, złożonego z siedmiu segmentów. Wyniki wyszukiwania można również zapisać – dotyczy to zarówno warstwy dźwiękowej, jak i warstwy transkrypcji.

Korpus tekstów i nagrań jest połączony ze słownikiem wyrazów, które wystąpiły w nagraniach i są nieznane w polszczyźnie ogólnej lub występują w gwarze w innym znaczeniu. Hasła wyrazów *stricte* gwarowych są podane w wersji standaryzowanej oraz w wersji gwarowej, np. *odziewaczka* [odziywacka] 'duża, gruba wełniana chusta, okrywająca głowę, ramiona i plecy, noszona przez starsze kobiety w czasie zimy'. Hasła wyrazów różniących się pod względem semantycznym mają postać standardową i gwarową, np. *wleźć* [wlyź] 'wejść'. Opracowanie zawiera leksemy, które notowane są w słownikach ogólnych jako dawne, przestarzałe i regionalne, np. *odziewać* [odziywać] 'zakładać ubranie; ubierać', *gazdówka* [gazdówka] 'gospodarstwo wiejskie na Spiszu i sąsiednich terenach'.

7. Zastosowanie

Językoznawstwo korpusowe to przede wszystkim pewna metodologia, która stwarza wiele możliwości, ale też stawia ograniczenia. Pierwsze ograniczenie to brak poświadczeń negatywnych: z faktu, że czegoś w korpusie nie znajdujemy, nie wynika, że jest językowo nieakceptowalne. Z tym problemem twórcy korpusu spotykali się wielokrotnie w odniesieniu do fleksji. Niewiele wyrazów ma potwierdzony pełny paradygmat fleksyjny. W oczywisty sposób bazą empiryczną dla fleksji musi być kompetencja rodzimego użytkownika języka.

7.1. Socjolingwistyka (zróżnicowanie języka) i geografia

Możliwość ograniczenia wyszukiwania za pomocą danych socjologicznych pozwala na szczegółową analizę wyników dla dowolnej grupy respondentów. Przykładowo można wyszukać wystąpienia słowa *dom* w wypowiedziach kobiet urodzonych przed rokiem 1950, mieszkających w Niedzicy. Tego rodzaju ograniczenia umożliwiają badanie wpływu na cechy językowe takich zmiennych demograficznych, jak wiek czy płeć[8].

7.2. Gramatyka: fleksja i składnia

Jak wyżej wspomniano, korpus jest zaopatrzony w tzw. anotację morfologiczną. Umożliwia to szereg badań nad gramatyką: fleksją, słowotwórstwem, a także składnią. Należy przypomnieć, że anotacja polega jedynie na opisaniu każdego słowa przez przysługujące mu kategorie gramatyczne, jednak nie ma tam jawnego podziału na morfemy; poszczególne morfemy można wyszukiwać, wpisując sekwencję liter[9], trzeba jednak pamiętać, że jest to jedynie przybliżenie. Podobnie przybliżeniem jest wyszukiwanie konstrukcji składniowych poprzez wyszukiwanie sekwencji wyrazów o danej charakterystyce fleksyjnej[10]. W korpusie tej wielkości częstsze zjawiska gramatyczne są już na tyle dobrze reprezentowane, że pozwala to na analizę ilościową i odróżnienie zjawisk marginalnych od typowych.

8 Na temat wpływu metadanych na kod respondentów powstaje osobny artykuł.

9 I tak np. zdrobnienia dają się wyszukać jako słowa składające się z dowolnej sekwencji liter, z których ostatnie to *-eczek*, *-eczka*, *-eczko*, a więc to, co wyszukuje zapytanie [lemma=".+*ecz(ek|ka|ko)*"].

10 Np. czas przyszły niedokonany to sekwencja czasownika *być* w czasie przyszłym i bezokolicznika lub formy przeszłej czasownika w dowolnej kolejności. Taki wzór nie zwróci jednak ciągów typu *będzie szybko szedł*.

7.3. Pragmatyka

Przypomnijmy, że *Korpus Spiski* jest zbiorem tekstów mówionych o rozmiarze porównywalnym z komponentem konwersacyjnym NKJP. Ponieważ wyszukiwarka korpusowa udostępnia podgląd szerszego kontekstu, może on służyć do badań nad pragmatyką. Taki szerszy kontekst jest niezbędny np. dla badania struktury tematyczno-rematycznej i jej wpływu na składnię. Dodajmy, że jest to problem wciąż nie do końca zbadany w odniesieniu do polszczyzny, również w wersji standardowej. Kolejne zagadnienie, które daje się rozwiązać przy pomocy korpusu, to badanie znaczników dyskursu (ang. *discourse markers*).

7.4. Fonetyka

Warstwa dźwiękowa jest znakomitym materiałem do badań nad fonetyką i prozodią. W tym celu od początku zdecydowano o bezstratnym formacie zapisu, nie zaś o zapisie w popularnym formacie MP3, którego niewątpliwą zaletą jest mniejszy rozmiar pliku – formaty bezstratne nadają się do badań instrumentalnych znacznie lepiej.

Każdy fragment wywiadu, który zawiera zazwyczaj ok. 15 sekund nagrania, można pobrać na własny komputer i przetwarzać za pomocą programu do analiz fonetycznych (np. Praat). Trzeba jednak zaznaczyć, że nie wszystkie nagrania są jakości takiej, która pozwalałaby na rzetelne badania fonetyczne. Priorytetem było zebranie możliwie dużego korpusu do wszechstronnych zastosowań i eksploratorzy nie rezygnowali z nagrania jedynie dlatego, że warunki akustyczne nie były dobre.

Warto dodać, że wyszukiwarka pozwala łatwo zebrać potrzebne dane, np. jedną z samogłosek w otoczeniu dwu spółgłosek zwartych. Co więcej, dzięki metadanym demograficznym można ograniczyć wyszukiwanie do informatorów o danej płci, określonym wieku i miejscu zamieszkania. Nagrania umożliwiają badanie prozodii, która jest istotnym elementem dociekań składniowych.

7.5. Leksyka

Choć przyjmuje się, że korpus tej wielkości nie jest dobrym źródłem leksyki, okazuje się, iż notuje on ok. 10 tys. wyrazów dyferencyjnych[11], co jest liczbą dosyć pokaźną. Zapewne taka obfitość słownictwa wynika również z tematyki rozmów, często dotyczącej zwyczajów, zabaw z dzieciństwa, pracy na roli itp. Ponadto tylko korpus umożliwia badanie kolokacji, wreszcie pozwala stworzyć słownik frekwencyjny.

11 Dla porównania *Słownik gwary orawskiej* (KąśSGO) zawiera ok. 28 tys. haseł, uwzględniając cały zasób leksyki, w tym także wspólnej z językiem ogólnym.

Na koniec dodajmy, że korpus ten jest również źródłem wiedzy o kulturze duchowej i obyczajach Spiszaków.

8. Podsumowanie

Żadnego z zarysowanych wyżej problemów nie da się rozwiązać bez dużego, zdygitalizowanego zbioru tekstów. Nie zastąpią go ani słownik, ani drukowane transkrypcje tekstów, ani wreszcie sam zbiór nagrań.

Trzeba też z całym naciskiem stwierdzić, że wielką wartością przedstawianego projektu jest właśnie jego objętość. Niewielki korpus daje na tyle rzadkie poświadczenia nawet stosunkowo częstych zjawisk, że pozwala jedynie na dość ograniczone badania. Środowisko dialektologiczne otrzymuje narzędzie, które co prawda punktowo, niemniej bardzo szczegółowo i wielostronnie pozwala zbadać jedną z gwar południowej Małopolski.

Literatura

AJPP: M. Małecki, K. Nitsch, 1934, *Atlas językowy polskiego Podkarpacia*, cz. I: *Mapy*, cz. II: *Wstęp, objaśnienia, wykazy wyrazów*, Kraków.

Brugman H., Russel A., 2004, *Annotating Multimedia/Multi-modal Resources with ELAN*, [w:] M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa, R. Silva (red.), *Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation*, Paris, s. 2065–2068, [on-line:] http://www.lrec-conf.org/proceedings/lrec2004/pdf/480.pdf (dostęp: 2 XI 2018).

Derwojedowa M., Kieraś W., Skowrońska D., Wołosz R., 2014, *Współczesne narzędzia leksykograficzne a analiza tekstów dawniejszych*, „Polonica" XXXIV, s. 21–27.

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

Garczyńska J. (red.), 2013–2017, *Akustyczna baza danych gwar mazowieckich. Wokalizm*, [on-line:] http://ec2-34-217-145-151.us-west-2.compute.amazonaws.com:8000/ (dostęp: 2 XI 2018).

Górski R.L., Łaziński M., 2012, *Reprezentatywność i zrównoważenie korpusu*, [w:] A. Przepiórkowski, M. Bańko, R.L. Górski, B. Lewandowska-Tomaszczyk (red.), *Narodowy Korpus Języka Polskiego*, Warszawa, s. 25–36.

Grochola-Szczepanek H., 2013, *Badanie języka mieszkańców wsi w kontekście przemian społecznych*, „Socjolingwistyka" XXVII, s. 43–53.

Grochola-Szczepanek H., 2017, *Nowe badania języka mieszkańców wsi regionu polskiego Spisza*, [w:] B. Osowski, P. Michalska-Górecka, J. Kobus, A. Piotrowska-Wojaczyk (red.), *Język w regionie, region w języku 2*, Poznań, s. 103–119.

Grochola-Szczepanek H., Woźniak M., 2018, *Transkrypcja języka mieszkańców wsi w aplikacji ELAN w Korpusie Spiskim*, [w:] R. Przybylska, M. Rak, A. Kwaśnicka-

-Janowicz (red.), *Historia języka, dialektologia i onomastyka w nowych kontekstach interpretacyjnych*, Kraków, s. 267–278.

Gruszczyński W., Adamiec D., Ogrodniczuk M., 2013, *Elektroniczny korpus tekstów polskich z XVII i XVIII w. (do 1772 r.) – prezentacja projektu badawczego*, „Polonica" XXXIII, s. 311–318.

Karaś H. (red.), 2010, *Dialekty i gwary polskie. Kompendium internetowe*, [on-line:] http://www.dialektologia.uw.edu.pl/index.php (dostęp: 2 XI 2018).

Karaś H., Kresa M., Krawczyk-Wieczorek A., 2012, *Towards a Corpus of Polish Dialect Texts*, „Prace Filologiczne" LXIII, s. 129–145.

KąśSGO: J. Kąś, *Słownik gwary orawskiej*, t. I–II, wyd. II, Kraków 2011.

KorBa: *Elektroniczny korpus tekstów polskich z XVII I XVIII w. (do 1772 r.)*, [on-line:] http://korba.edu.pl/ (dostęp: 4 II 2019).

Krawczyk-Wieczorek A., 2012, *Automatyczna lematyzacja tekstu w zapisie fonetycznym. Korpus polskiej gwary południowokresowej*, „Język Polski" XCII, s. 11–19.

Lubaś W., 1979, *Społeczne uwarunkowania współczesnej polszczyzny. Szkice socjolingwistyczne*, Kraków.

Narodowy Korpus Języka Polskiego, praca zbiorowa, red. A. Przepiórkowski, M. Bańko, R.L. Górski, B. Lewandowska-Tomaszczyk, Warszawa 2012.

NKJP: Narodowy Korpus Języka Polskiego, [on-line:] nkjp.pl.

Przepiórkowski A., 2004, *Korpus IPI PAN. Wersja wstępna*, Warszawa.

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.

Wyderka B., 2014, *Problemy teoretyczne współczesnej dialektologii*, [w:] M. Rak, K. Sikora (red.), *Badania dialektologiczne. Stan, perspektywy, metodologia*, „Biblioteka LingVariów", t. 17, Kraków, s. 13–21.

A Spoken Corpus of Inhabitants of Polish Spisz
Summary

The article describes a dialect corpus project that documents the dialect of Polish Spisz. In contrast to the majority of dialectological research in Poland, our corpus also includes the speech of the youngest and middle generations, as its aim is also to document the sociolinguistic situation of the dialect of the region. Recordings have been transcribed into standard Polish orthography, not phonetically, which makes it possible not only to easily search the corpus but also to use existing tools to lemmatize and add morphosyntactic annotation to the texts. Users interested in the phonetic layer can access the recordings on a per-utterance basis. The article describes the stages of compiling the corpus and discusses its potential applications. The authors argue that a large corpus which covers a small, homogeneous area is a more valuable resource for dialectologists than a series of small corpora documenting a larger region.