LingVaria
IX (2014) | 2 (18)
doi:10.12797/LV.09.2014.18.14

Justyna Garczyńska
Uniwersytet Warszawski, Warszawa
jgarczynska@uw.edu.pl

PROJEKT AKUSTYCZNEJ BAZY DANYCH GWAR MAZOWIECKICH

Słowa klucze: językoznawstwo, dialektologia, fonetyka akustyczna, dialekt mazowiecki, samogłoska
Keywords: linguistics, dialectology, acoustic phonetics, Mazovian dialect, vowel

1. Cel projektu[1]

Celem projektu „Akustyczna baza danych gwar mazowieckich. Wokalizm" jest opracowanie ogólnodostępnej i reprezentatywnej statystycznie akustycznej bazy danych szeroko rozumianego dialektu mazowieckiego. Baza będzie zawierać obiektywne, niezależne od odsłuchu badacza dane w postaci wartości docelowych dwóch pierwszych formantów dla akcentowanych i nieakcentowanych samogłosek ustnych w różnych kontekstach spółgłoskowych. Planowana baza zostanie uzupełniona o odpowiednie dane akustyczne dotyczące samogłosek polszczyzny ogólnej. Zawarte w niej dane umożliwią automatyczne porównywanie gwarowych realizacji samogłosek z realizacjami odpowiednich głosek polskiego języka ogólnego, ukazując współczesny obraz wokalizmu gwar mazowieckich i pozwalając na weryfikację dotychczasowych ustaleń. Baza obejmie średnio 60 000 realizacji samogłosek, zatem będzie reprezentatywna statystycznie. Będzie też opatrzona w bogate metadane, takie jak: dialekt, gwara, miejsce urodzenia informatorów, ich wiek, płeć, wykształcenie, data i źródło

1 Projekt realizowany przez zespół złożony z czterech osób w ramach grantu naukowo-badawczego Narodowego Centrum Nauki pt. „Akustyczna baza danych gwar mazowieckich. Wokalizm".

nagrania. Ponadto zostanie wyposażona w wyszukiwarkę pozwalającą formułować skomplikowane kwerendy, a także w bibliotekę wykresów i danych statystycznych charakteryzujących wymowę informatorów.

Baza danych w założeniu nie będzie korpusem tekstów mazowieckich odzwierciedlających wszystkie cechy gwarowe danego regionu. Celem projektu jest analiza akustyczna samogłosek, która będzie stanowić punkt wyjścia późniejszych opracowań. Stąd nie przewiduje się pełnego zapisu fonetycznego tekstów, który jest już w pewnym stopniu interpretacją materiału dźwiękowego, lecz zapis powszechnie stosowanym w bazach akustycznych alfabetem SAMPA, który nie odzwierciedla, np. stopnia ścieśnienia samogłosek. Tego typu zjawiska samogłoskowe użytkownik bazy będzie mógł sam zinterpretować na podstawie załączonych danych i wykresów. Baza będzie miała zatem charakter uniwersalny jako zbiór wysokiej jakości nagrań i obiektywnych danych dotyczących samogłosek, stanowiący punkt wyjścia dla dalszych badań. Niemniej zostaną w niej uwzględnione najważniejsze cechy wymowy spółgłosek charakterystyczne dla dialektu mazowieckiego, takie jak mazurzenie, np. *capka* ‘czapka’, sziakanie, np. *cziapka* ‘czapka’, niekonsekwentna wymowa grup *ke-kie*, *ge-gie*, np. *rękie*, *cuker*, dźwięczna wymowa *h* i asynchroniczna realizacja spółgłosek wargowych miękkich, np. *bziały* ‘biały’.

W pracach nad opisywaną bazą po raz pierwszy w polskiej dialektologii na tak dużą skalę zostaną zastosowane metody fonetyki akustycznej, a także statystyka. Metody akustyczne pozwalają w pewnym stopniu na rozstrzygnięcie podstawowego dla dialektologa zagadnienia, jakim jest stosunek percepcji słuchowej do rzeczywistego brzmienia głosek, natomiast dzięki statystyce możliwe jest wykrycie artykulacji typowych oraz istniejących w mowie informatorów tendencji fonetycznych.

2. Znaczenie projektu

2.1. Baza danych akustycznych jako źródło wyczerpujących i aktualnych danych fonetycznych z terenu całego Mazowsza

Mazowsze jako całość jest pod względem fonetycznym dzielnicą właściwie nieopracowaną. Dotychczasowe prace z tego zakresu dotyczą zazwyczaj gwary jednego regionu lub wybranego zjawiska fonetycznego, a wymowa samogłosek jest w nich traktowana dość pobieżnie. Jedyna istniejąca monografia fonetyczna z zakresu wokalizmu dialektu mazowieckiego – *Studia nad wokalizmem w gwarach Mazowsza* (*samogłoski ustne*) Anny Basary (1965), obejmuje tylko teren Mazowsza właściwego i opiera się na faktach językowych zbieranych w latach 50. XX w., jest już zatem w dużym stopniu nieaktualna. Od tego czasu nie były prowadzone kompleksowe badania fonetyczne nad dialektem mazowieckim. Planowana akustyczna baza danych wypełni zatem lukę badawczą zarówno pod względem obszaru badań, jak i aktualności danych fonetycznych.

Od 2007 r. w Instytucie Języka Polskiego Wydziału Polonistyki UW są prowadzone badania gwaroznawcze na całym terenie Polski. W obrębie dialektu mazowieckiego zebrano teksty z ponad pięćdziesięciu wsi. Część z nich zostanie wykorzystana w projekcie. Planowane są dalsze nagrania na terenie całego szeroko rozumianego dialektu mazowieckiego (10 regionów)[2], czyli Mazowsza bliższego, Mazowsza dalszego, Kurpiów, Mazur, a przede wszystkim na obszarach Warmii, Ostródzkiego, Lubawskiego, Podlasia, Suwalszczyzny i Łowickiego, które są słabiej reprezentowane pod względem nagranych tekstów. Planowane jest zbadanie co najmniej sześciu wsi z każdego regionu oraz jednego informatora z każdej wsi. W badaniach będą uczestniczyć kobiety i mężczyźni w wieku od 20 do 80 lat, mieszkający od urodzenia na terenie danego regionu. W dotychczasowych badaniach fonetycznych dialektu mazowieckiego najczęściej brały udział osoby w wieku powyżej 50 lat. Uwzględnienie w planowanym projekcie także osób młodszych w pełni ukaże zróżnicowanie fonetyczne oraz pozwoli na nakreślenie tendencji wymawianiowych w badanych gwarach w obrębie samogłosek. Otrzymane z wykorzystaniem tej samej metodologii dane będą porównywalne ze sobą, dając spójny i aktualny obraz wokalizmu gwar mazowieckich, który pozwoli na weryfikację dotychczasowych ustaleń tak szczegółowych zagadnień, jak podwyższona wymowa samogłoski [y], charakter samogłosek pochylonych, przednia artykulacja samogłoski [a], wpływ akcentu na artykulację samogłosek, wpływ kontekstu miękkiego i sonornego na realizację samogłosek, różnice między mową kobiet i mężczyzn w zakresie wokalizmu, różnice wymawianiowe zależne od wieku lub wykształcenia respondentów.

2.2. Baza danych akustycznych jako prezentacja nowej metodologii w badaniach dialektologicznych

Projekt wykorzystuje w szerokim stopniu osiągnięcia metodologiczne fonetyki akustycznej i statystyki. Takie podejście do analizowanego materiału wymogła obserwowana w gwarach silna wariantywność fonetyczna, wynikająca z działania rozmaitych czynników fonetycznych, socjalnych i indywidualnych, dotyczących badanych informatorów, takich jak miejsce pochodzenia, wiek, wykształcenie, wpływ języka ogólnopolskiego czy kontekst, w jakim znalazła się badana głoska. Wariantywność ta często jest nieuchwytna dla ludzkiego ucha i jej precyzyjny opis wymaga zastosowania specjalistycznego programu komputerowego do akustycznej analizy dźwięku.

Podstawowym zagadnieniem dla dialektologa jest stosunek percepcji słuchowej do rzeczywistego brzmienia głosek. Do tej pory stosuje się trzy metody odczytywania tekstów gwarowych utrwalonych fonograficznie: odsłuch indywidualny, odsłuch zespołowy oraz odsłuch indywidualny lub zespołowy poparty analizą akustyczną.

2 Dialekt mazowiecki obejmuje Mazowsze właściwe i Mazury, a także gwary podlaskie i suwalskie (Urbańczyk 1968, m. 3). W szerokim ujęciu z pewnymi zastrzeżeniami do dialektu mazowieckiego zalicza się także Warmię, Ostródzkie i Lubawskie (EJP, m. 1), a w części prac również Łowickie (Kowalska 1991: 128).

Pierwsza z metod, najszerzej wykorzystywana, jest metodą subiektywną, i co za tym idzie, niezbyt dokładną, istnieje bowiem szereg subiektywnych czynników, które wpływają na różnice w zapisach dokonanych na podstawie percepcji słuchowej tego samego nagranego tekstu przez kilku dialektologów lub między kilkukrotnymi zapisami tej samej osoby. Rozrzut, czyli rozchwianie percepcji słuchowej, przy założeniu, że każdy z transkrybentów ma słuch w granicach normy, zależy od następujących czynników: przesłyszenia się, indywidualnych tendencji do słyszenia pewnych zjawisk fonetycznych w określony sposób, przypadkowej niedyspozycji transkrybenta, autosugestii polegającej na tym, że dialektolog może słyszeć nie to, co się wymawia faktycznie, ale to, czego oczekuje, oraz od rzeczywistego rozchwiania artykulacyjnego sygnałów mowy u nadawcy (Sobierajski, Nowak, Gruchmanowa 1962: 11–18). W wypadku samogłosek problem pogłębia to, że głoski tego typu, w przeciwieństwie do spółgłosek, nie mają stałego, określonego miejsca artykulacji, czego rezultatem jest pewien rozrzut konfiguracji toru ustno-gardłowego podczas ich wymawiania. Możliwość w miarę dokładnego ustalenia pierwszych czterech subiektywnych czynników i w związku z tym określenia rzeczywistego stopnia rozchwiania wymowy w danej gwarze to zaleta drugiej z metod, mianowicie metody zespołowego odsłuchu. Najlepszą jednak z wymienionych metod jest trzecia, w dialektologii polskiej niestosowana, w której słuch transkrybenta wspierany jest analizą fali głosowej wytwarzanej podczas artykulacji głosek przez nadawcę. Choć przez kilkadziesiąt ostatnich lat akustyczna analiza mowy poczyniła szybkie postępy, prace dotyczące zjawisk fonetycznych zachodzących w polskich gwarach w dalszym ciągu są prowadzone na podstawie odsłuchu indywidualnego z nie najlepszej jakości nagrań, co sprawia, że są obciążone dużym ryzykiem błędu.

Wstępne badania prowadzone metodami akustycznymi nad samogłoskami występującymi w gwarach polskich w kraju i poza jego granicami ukazują ogromne możliwości proponowanej w projekcie metodologii. Zastosowanie do opisu ustnych samogłosek metod wypracowanych przez fonetykę akustyczną pozwala na:

a) ustalenie, w jaki sposób jest zorganizowana struktura wewnętrzna zbiorów pomiarów F1 i F2, czyli jaki jest stopień zróżnicowania wymowy badanych samogłosek określony współczynnikiem rozproszenia V, jakie są przyczyny tego zróżnicowania (kontekstowe, związane z tempem mowy, wpływem polszczyzny ogólnej czy istnieniem reliktów dawnych samogłosek pochylonych);
b) odtworzenie mechanizmu artykulacyjnego badanych samogłosek, co umożliwia istnienie korelacji między wartościami częstotliwości F1 i F2 a pozycjami artykulacyjnymi w ponadkrtaniowym kanale głosowym;
c) porównanie samogłoskowych czworoboków artykulacyjnych uzyskanych dla poszczególnych osób z odpowiednimi czworobokami dla samogłosek polszczyzny ogólnej;
d) określenie stopnia wpływów języka ogólnopolskiego i innych języków w mowie badanych osób.

Zastosowanie w planowanej bazie danych metod fonetyki akustycznej jest zatem pomysłem prekursorskim w polskiej dialektologii. Dane, których dostarcza fonetyka akustyczna, pozwalają na obiektywne, precyzyjne i wyczerpujące charakteryzowanie dźwięków mowy. Komputer umożliwia szybkie uzyskiwanie spektrogramów głosek występujących w mowie informatorów oraz łatwe porównywanie ich między sobą i z danymi akustycznymi dla samogłosek innych języków. Ponadto technizacja warsztatu badawczego dialektologa fonetyka jest konieczna, jeśli wyniki jego badań mają dorównywać poziomowi innych nauk, przede wszystkim ścisłych, przyrodniczych i technicznych.

3. Metodologia

Podstawę warsztatu naukowego projektu stanowią przede wszystkim metody fonetyki akustycznej i metody przyjmowane w dialektologii (np. w geografii lingwistycznej w zakresie doboru tekstów).

Słuchowa identyfikacja samogłosek zależy od ich budowy akustycznej. Porównując widma różnych samogłosek, można stwierdzić, że najwyraźniej różnią się one wartościami częstotliwości formantów, czyli częstotliwościami rezonansowymi toru głosowego. Na podstawie badań nad polskimi samogłoskami izolowanymi, a także występującymi w mowie ciągłej stwierdzono, że dla ich rozpoznawania wystarczające są dwa pierwsze formanty (Jassem, Krzyśko, Dyczkowski 1972), stąd też akustyczny opis samogłosek ogranicza się często do podania częstotliwości tych formantów. Formanty trzeci i czwarty zawierają cechy indywidualne nadawcy. Istnieją ściśle określone związki między częstotliwościami formantów F1 i F2 a cechami artykulacyjnymi samogłosek. Częstotliwość F1 jest tym wyższa, im większy jest stopień otwarcia jamy ustnej, zatem F1 rośnie od [i], poprzez [y], [e] do [a], a następnie opada poprzez [o] do [u]. Natomiast częstotliwość F2 podnosi się w miarę przesuwania się miejsca najwyższego wzniesienia języka w kierunku wylotu ustnego, zatem F2 konsekwentnie rośnie w kolejności [u], [o], [a], [e], [y], [i] (Jassem 1973: 191–193). Opisane wyżej zależności akustyczno-artykulacyjne można zademonstrować, przedstawiając wartości częstotliwości formantów różnych samogłosek polskich w polu dwuwymiarowym, którego oś X przyporządkowuje dźwięki samogłoskowe pod względem pierwszego, a oś Y – pod względem drugiego formantu. Uzyskuje się w ten sposób duże podobieństwo do artykulacyjnego trójkąta – czworoboku artykulacyjnego (zob. rys.1).

Rodowici użytkownicy danego języka najczęściej odbierają samogłoski wypowiedziane na przykład w kolejnych powtórzeniach tego samego wyrazu jako takie same. W rzeczywistości każda samogłoska, słyszana jako identyczna, ma strukturę akustyczną (wartości F1; F2) mniej lub bardziej różną od pozostałych. Nawet pomiary częstotliwości F1; F2 wykonane dla identycznych słuchowo samogłosek, wymówionych w tych samych warunkach przez tę samą osobę, wykazują pewien rozrzut wartości (Dukiewicz, Sawicka 1995: 62). Może to wynikać z wyżej wymienionych czynników, a także z braku precyzji

w realizacji samogłosek, które nie mają wyraźnie określonego miejsca artykulacji. Ta cecha samogłosek powoduje pewien rozrzut konfiguracji toru ustno-gardłowego, a co za tym idzie, rozrzut wartości F1; F2. Realizacja danej samogłoski na wykresie rozrzutu nie wygląda zatem jak pojedynczy punkt, ale jak chmura punktów (zob. rys.2).

Widoczna na wykresie wariantywność samogłosek zwiększa się w mowie gwarowej, gdy dochodzą takie czynniki różnicujące, jak pochodzenie, wykształcenie lub wiek informatora. To zróżnicowanie wymowy samogłosek gwarowych, nieuchwytne ludzkim uchem, daje się zbadać i opisać dzięki metodologii fonetyki akustycznej.

3.1. Analiza porównawcza

Badania dialektologiczne obejmują także prowadzenie analiz porównawczych. Mogą one mieć na celu określenie, w jaki sposób badane samogłoski różnią się pod względem akustyczno-artykulacyjnym od odpowiednich głosek języka polskiego ogólnego, w jakim stopniu języki współwystępujące na danym terytorium wpłynęły na ich wymowę, a także jak różni się wymowa danej głoski u poszczególnych respondentów w poszczególnych regionach. Do utworzenia normy porównawczej języka ogólnopolskiego posłużyły odpowiednio dobrane zdania i wyrazy nagrane wśród pracowników i studentów Uniwersytetu Warszawskiego. W wypadku samogłosek różnice między badanymi głoskami i głoskami uwzględnionych w analizie porównawczej języków można pokazać przez:

a) naniesienie na płaszczyznę o współrzędnych F1 i F2 zakresu wartości dwóch pierwszych formantów dla danej głoski wymówionej przez daną osobę i tychże wartości dla odpowiednich samogłosek porównywanych języków;
b) przedstawienie na jednej płaszczyźnie czworoboków artykulacyjnych uzyskanych dla badanych samogłosek i samogłosek porównywanych języków.

W analizach wykorzystujących metody akustyczne porównawczą bazę danych należy budować z dużą ostrożnością, ponieważ istnieje szereg czynników zmieniających wartości F1, F2, np. płeć, wiek mówiącego, to, czy samogłoska jest izolowana, czy w kontekście itp. Oczywiste zatem jest, że porównywane dane muszą sobie odpowiadać pod względem wymienionych czynników. W planowanej akustycznej bazie kryterium odpowiedniości materiału gwarowego i porównawczych danych języka ogólnopolskiego zostało spełnione.

3.2. Dobór próby badawczej

Próbę badawczą, służącą do przeprowadzenia analiz akustycznych, będą stanowiły samogłoski wyekscerpowane z nagranych tekstów ze wszystkich regionów dialektu

mazowieckiego. Długość nagranych tekstów będzie wynosić około godziny, a tematyka będzie związana przede wszystkim z życiem na wsi (w wypadku starszych respondentów) lub będzie dowolna (przy młodszym pokoleniu). Do analizy zostaną wybrane wyrazy, w których poszczególne samogłoski znajdą się:

a) w obustronnym sąsiedztwie spółgłosek twardych;
b) w obustronnym lub jednostronnym kontekście spółgłosek miękkich;
c) w obustronnym lub jednostronnym kontekście spółgłosek sonornych;
d) w wygłosie po spółgłoskach twardych.

Planuje się uzyskanie co najmniej trzydziestu przykładów dla każdej akcentowanej i nieakcentowanej samogłoski w wymienionych kontekstach wymówionej przez każdego z informatorów. Jest to liczba niezbędna dla uzyskania statystycznej reprezentatywności badanego materiału.

3.3. Nagrania i pomiary

Do nagrań planuje się wykorzystać dyktafony Olympus WS-700M, które dają możliwość prowadzenia nagrań z częstotliwością 44 kHz w formacie bez kompresji Linear PCM. Pomiary częstotliwości formantowych samogłosek będą przeprowadzane z wykorzystaniem programu do cyfrowej analizy sygnału mowy Praat przygotowanego przez Paula Boersme i Davida Weeninka (*Phonetic Science Departament at the University of Amsterdam*). Wprowadzony do komputera materiał zostanie poddany segmentacji i analizie formantowej. W analizie uwzględniono częstotliwości formantu pierwszego i drugiego, które najwyraźniej są związane z artykulacyjnymi przesunięciami w ponadkrtaniowym kanale głosowym. Dla każdej z analizowanych samogłosek będzie brana pod uwagę częstotliwość docelowa, odczytywana w stadium ustalonym dźwięku, najczęściej przypadająca w środkowej części głoski. Parametry F1 i F2 będą stanowiły wektor cech dla dalszych obliczeń statystycznych.

3.4. Transkrypcja fonetyczna

Opis sygnału mowy wymaga nadania etykiet poszczególnym jego segmentom. Wygodny w przetwarzaniu komputerowym i powszechnie stosowany w akustycznych bazach danych jest alfabet SAMPA (*Speech Assessment Methods Phonetic Alphabet*) i ten sposób zapisu zostanie zastosowany. Pozwoli to na korzystanie z bazy językoznawcom z innych niż slawistyczne ośrodków naukowych, a jednocześnie umożliwi zachowanie najważniejszych mazowieckich fonetycznych cech gwarowych.

4. Koncepcja i plan badań

1. Opracowanie instrukcji, uporządkowanie materiału dźwiękowego zgromadzonego w wyniku wcześniejszych badań, utworzenie siatki zbadanych punktów i uzupełnienie jej o nowe punkty badawcze.

W pierwszym, wstępnym etapie projektu zostanie opracowana instrukcja dotycząca:

a) sposobu nagrywania i doboru informatorów oraz tematyki nagrywanych tekstów gwarowych;
b) sposobu nagrywania i doboru tekstów oraz informatorów w celu utworzenia ogólnopolskiej normy wymawianiowej samogłosek;
c) sposobu wydzielania mniejszych segmentów (zdań, fraz) oraz wyrazów i samogłosek z tekstów (wybór programu do segmentacji tekstów, ustalenie maksymalnej długości segmentu, ustalenie miejsca odczytu wartości formantów samogłoskowych);
d) sposobu oznaczania materiału (tekstów, zdań/fraz, wyrazów, kontekstów samogłosek, samogłosek);
e) sposobu przechowywania danych (płyty CD, twardy dysk komputera, serwer UW, dyski przenośne);
f) zestawu metadanych (przewidywane są następujące pola metadanych: obszar gwarowy (region), wieś i jej współrzędne geograficzne, powiat, województwo, parafia (istotna informacja na terenach mieszanych wyznaniowo), data przeprowadzenia nagrania, data i miejsce urodzenia informatora, miejsce urodzenia rodziców informatora, miejsce urodzenia dziadków informatora, wykształcenie informatora, wykształcenie rodziców informatora, wykształcenie dziadków informatora, opis warunków technicznych nagrań);
g) zakresu i sposobu działania wyszukiwarki danych na stronie internetowej akustycznej bazy danych gwar polskich.

Następny etap prac zakłada uporządkowanie istniejącego już (przechowywanego w IJP UW) zbioru nagranych tekstów z obszaru dialektu mazowieckiego. W pierwszej kolejności z dalszych badań zostaną wyłączone wszystkie nagrania mające niepełne metadane, których brak uniemożliwia wyczerpującą analizę samogłosek, np. niezawierających informacji o wieku i miejscu urodzenia informatora oraz nagrania w formatach skompresowanych (mp3, mp4, wma itp.). Z pozostałej grupy nagrań do planowanej bazy zostaną wybrane teksty spełniające kryteria geolingwistyczne (pochodzące z odpowiedniego regionu) oraz socjolingwistyczne (przede wszystkim kryterium wieku informatora). Na podstawie wybranych tekstów mazowieckich zostanie utworzona mapa zbadanych już wsi, która zostanie następnie uzupełniona o miejsca do dalszych terenowych badań gwaroznawczych.

2. Nagrania tekstów polszczyzny ogólnej (zdań i wyrazów) w celu utworzenia ogólnopolskiej normy wymawianiowej samogłosek.

Ten etap będzie polegał na przeprowadzeniu nagrań w grupie respondentów obejmującej trzy kategorie wiekowe (20–40; 40–60; 60–80 lat), mających wykształcenie wyższe, posługujących się na co dzień polszczyzną ogólną i niewywodzących się ze środowisk wiejskich, ale zamieszkujących obszar Mazowsza w celu wykluczenia ewentualnych różnic regionalnych. Grupę docelową badań będą stanowić pracownicy i studenci Wydziału Polonistyki UW.

3. Opracowanie projektu strony internetowej akustycznej bazy danych oraz internetowej wyszukiwarki danych.

Na tym etapie prac zostanie opracowany projekt strony internetowej oraz wyszukiwarki danych obejmującej dane z bazy nagranych tekstów (dialektalnej i ogólnopolskiej), bazy wyekscerpowanych zdań/fraz, bazy wyrazów, bazy kontekstów spółgłoskowych, bazy wartości dwóch pierwszych formantów samogłosek F1;F2 oraz bazy metadanych. Dane będzie można przeszukiwać między innymi przez wpisanie w polu wyszukiwarki metadanych lub kontekstów samogłosek. Na stronie internetowej zostaną zamieszczone także wszystkie istotne informacje o planowanej bazie oraz o procesie jej tworzenia, m.in. o założeniach, tekstach zawartych w korpusie i ich pochodzeniu, przyjętych zasadach transkrypcji, a także instrukcje wyszukiwania.

4. Przesłuchanie i podział tekstów mazowieckich na mniejsze całości oraz stworzenie bazy obejmującej nagrane teksty i metadane.

Na tym etapie nagrane teksty zostaną uporządkowane, oznakowane i sprawdzone pod względem kompletności metadanych, a następnie podzielone na mniejsze całości tematyczne (np. o uprawie roli, o hodowli pszczół) oraz zdania/frazy. Podzielone teksty zostaną zapisane alfabetem SAMPA. Zakłada się również uruchomienie na serwerze UW strony internetowej zawierającej informacje o projekcie oraz uporządkowane teksty wraz z podstawową wyszukiwarką dającą możliwość wyszukiwania nagrań według podanych metadanych.

5. Ekscerpcja wyrazów zawierających samogłoski w różnych kontekstach spółgłoskowych z wydzielonych zdań/fraz.

Wydzielone zdania/frazy zostaną podzielone na wyrazy, w których samogłoski znajdują się w różnych pozycjach (pod akcentem, nieakcentowane, w wygłosie) i w otoczeniu różnych grup spółgłosek (twardych, miękkich, sonornych).

6. Ekscerpcja samogłosek w różnych pozycjach i kontekstach z wyodrębnionych wyrazów.

Ten etap prac zakłada wyodrębnienie z wydzielonych wcześniej wyrazów samogłosek akcentowanych i nieakcentowanych w różnych kontekstach spółgłoskowych. Zakłada się wyodrębnienie następujących kontekstów spółgłoskowych: twardego, np. ***bab****a*, obustronnie miękkiego, np. ***ciocia***, lewostronnie miękkiego, np. ***ciast****o*, prawostronnie miękkiego, np. ***pasi****e*, obustronnie sonornego, np. ***lal****a*, lewostronnie sonornego, np. ***rak***, prawostronnie sonornego, np. ***pal***. Uwzględnione zostaną także samogłoski w wygłosie występujące po spółgłoskach twardych.

7. Utworzenie bazy danych wyodrębnionych samogłosek wraz z kontekstami.

Na tym etapie prac zostaną uporządkowane, oznakowane i uzupełnione o odpowiednie metadane wyodrębnione konteksty wraz z samogłoskami. Planowane jest także dołączenie powstałej bazy kontekstów do istniejącej już strony internetowej i uzupełnienie wyszukiwarki o nowe funkcje. Zakłada się, że po wpisaniu w polu wyszukiwarki, np. **BaB**, ukażą się wszystkie wyrazy, w których ten kontekst wystąpił wraz z odpowiednimi metadanymi.

8. Przesłuchanie i podział tekstów języka ogólnopolskiego na zdania i wyrazy oraz stworzenie bazy obejmującej nagrane zdania, wyrazy i metadane.

Ten etap prac obejmuje podział nagrań języka ogólnopolskiego na zdania i wyrazy oraz ich zapis alfabetem SAMPA. Zakłada się utworzenie bazy nagranych wyrazów i zdań wraz z zestawem metadanych i jej dołączenie do istniejącej strony internetowej.

9. Ekscerpcja samogłosek w różnych pozycjach i kontekstach z wyodrębnionych wyrazów języka ogólnopolskiego.

Ten etap prac zakłada wyodrębnienie z wydzielonych wcześniej wyrazów polszczyzny ogólnej samogłosek akcentowanych i nieakcentowanych w różnych kontekstach spółgłoskowych, podobnie jak to zostało opisane w punkcie 6.

10. Utworzenie bazy danych wyodrębnionych samogłosek języka ogólnopolskiego wraz z kontekstami.

Na tym etapie prac zostaną uporządkowane, oznakowane i uzupełnione o odpowiednie metadane wyodrębnione konteksty wraz z samogłoskami. Planowane jest także dołączenie powstałej bazy kontekstów do istniejącej już strony internetowej i uzupełnienie wyszukiwarki o nowe funkcje, podobnie jak to zostało opisane w punkcie 7.

11. Analiza akustyczna w programie Praat samogłosek w różnych pozycjach i kontekstach.

Na tym etapie zostanie przeprowadzona akustyczna analiza gwarowych mazowieckich i ogólnopolskich samogłosek ustnych we wszystkich uwzględnionych w projekcie pozycjach i kontekstach. Celem analizy jest uzyskanie wartości docelowych dwóch pierwszych formantów F1; F2.

12. Utworzenie biblioteki wykresów rozrzutu realizacji badanych samogłosek dla poszczególnych informatorów na tle norm ogólnopolskich.

Ten etap prac zakłada utworzenie dla każdego informatora dokładnego obrazu zróżnicowania wymowy pojedynczej samogłoski na tle normy ogólnopolskiej, np. akcentowanej samogłoski [a] w kontekście lewostronnie miękkim. Użytkownik akustycznej bazy danych gwar mazowieckich będzie zatem mógł nie tylko dowiedzieć się, jakie są wartości formantów dla danej samogłoski i danego informatora, ale także „obejrzeć" wymowę informatora na przejrzystym wykresie.

13. Utworzenie biblioteki wykresów czworoboków artykulacyjnych samogłosek dla poszczególnych informatorów na tle norm ogólnopolskich.

Ten etap prac zakłada utworzenie dla każdego informatora czworoboków artykulacyjnych wymawianych przez niego wszystkich samogłosek w określonym kontekście, np. samogłosek akcentowanych w kontekście twardym na tle normy ogólnopolskiej.

14. Utworzenie biblioteki tabel zawierających podstawowe statystyki dla poszczególnych samogłosek i informatorów.

Na tym etapie projektu zakłada się utworzenie tabel dla każdej samogłoski w mowie każdego informatora, zawierających statystyki podstawowe, takie jak liczba wymówień, średnia wartość F_1; F_2, zakres wartości F_1; F_2, odchylenie standardowe oraz współczynnik rozproszenia V.

5. Spodziewane rezultaty

Akustyczna baza danych gwar polskich będzie miała ogromne znaczenie zarówno naukowe, jak i dydaktyczne. Będzie stanowiła unikatowy w polskim językoznawstwie zbiór tekstów oraz gwarowych danych akustycznych, które mogą stanowić punkt wyjścia do dalszych badań, a także być wykorzystywane na zajęciach uniwersyteckich. Nie do przecenienia jest również wartość dokumentacyjna planowanej bazy dzięki utrwaleniu w postaci nagrań gwar zanikających lub szybko ewoluujących w wyniku przeobrażeń cywilizacyjnych.

Zastosowanie do opisu fonetycznego samogłosek ustnych metod wypracowanych przez fonetykę akustyczną i stworzenie akustycznej bazy danych w znaczący sposób uzupełni dotychczasową wiedzę z zakresu takich dziedzin, jak:

a) fonetyka akustyczna – przez uzupełnienie badań akustycznych nad polskim językiem ogólnym danymi dotyczącymi gwar;
b) fonetyka opisowa języka polskiego – przez ukazanie zróżnicowania wymowy badanych głosek, zależnego między innymi od pozycji w wyrazie i kontekstu

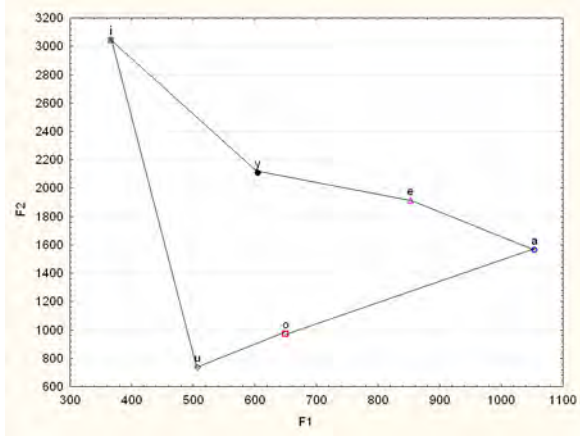
spółgłoskowego, które przyniesie lepsze zrozumienie mechanizmów koartykulacji, wpływu akcentu czy tempa mowy na artykulację samogłosek;
c) historia języka polskiego – przez określenie struktury akustycznej samogłosek pochylonych i charakteru zjawiska pochylenia;
d) językoznawstwo porównawcze – ponieważ uzyskanie na podstawie danych akustycznych czworoboków samogłoskowych dla gwar mazowieckich umożliwi ich porównanie z odpowiednimi czworobokami dla samogłosek polszczyzny ogólnej, a także dla innych gwar i języków;
e) dialektologia języka polskiego – przez uzyskanie aktualnych danych dotyczących fonetyki gwar mazowieckich i weryfikację dotychczasowych ustaleń;
f) zasoby planowanej bazy mogą być przydatne także dla informatyków zainteresowanych przetwarzaniem języka naturalnego oraz dla osób zajmujących się analizą mowy na potrzeby kryminalistyki.

Literatura

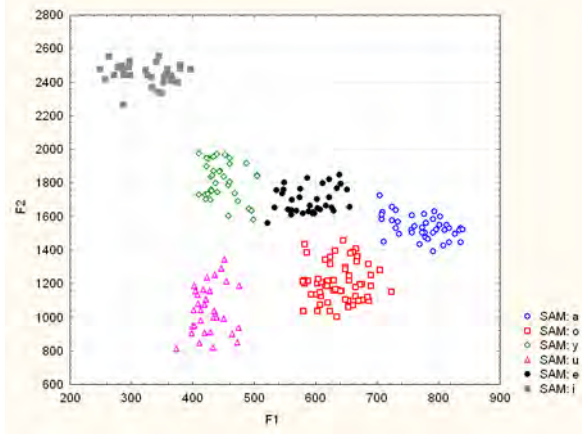
Dukiewicz L., Sawicka I., 1995, *Fonetyka i fonologia*, Kraków.
EJP: S. Urbańczyk i M. Kucała (red.), *Encyklopedia języka polskiego*, Wrocław 1999.
Jassem W., 1973, *Podstawy fonetyki akustycznej*, Warszawa.
Jassem W., Krzyśko M., Dyczkowski A., 1972, *Klasyfikacja i identyfikacja samogłosek polskich na podstawie częstotliwości formantów*, „Prace Instytutu Podstawowych Problemów Techniki" nr 64, Warszawa.
Kowalska A., 1991, *Podziały językowe Mazowsza na tle podziałów pozajęzykowych*, Warszawa.
Sobierajski Z., Nowak H., Gruchmanowa M., 1962, *Zastosowanie odsłuchu zespołowego do odczytywania tekstów gwarowych z płyt gramofonowych*, „Biuletyn Fonograficzny" V, Poznań, s. 11–41.
Urbańczyk S., 1968, *Zarys dialektologii polskiej*, Warszawa.

A project of an acoustich data base of Mazovian dialects
Summary

The paper presents a project of an acoustic data base of Mazovian dialects, and shows how modern research methods of acoustic phonetics can be employed in dialectological practice. The projected data base is going to contain objective data, independent of the researcher's hearing, in the form of target values of the first two formants for both stressed and unstressed oral vowels in various consonant contexts. The base is going to contain about 60,000 realizations. It is also going to be supplied with rich metadata, such as dialect, subdialect, place of birth of the speakers, their age, gender and education, the date and source of the recording. The base is going to be accessed through a search engine capable of executing complex queries, and it is going to contain a library of graphs and statistical data illustrating the informants' pronunciation. These data will allow for automated comparisons to be performed between the dialectal and the general Polish realizations of vowels, thus revealing the contemporary state of the vocalism of Mazovian dialects, and making it possible to verify the previous findings.



Rysunek 1. Odległości akustyczne między samogłoskami [i, y, e, a, o, u], wymówionymi w izolacji F1



Rysunek 2. Pola formantowe fonemów /i, y, e, a, o,u/