Studia z Filologii Polskiej
i Słowiańskiej, 49
SOW, Warszawa 2014

DOI: 10.11649/sfps.2014.013

Dorota Krystyna Rembiszewska
(Instytut Slawistyki PAN, Warszawa)

Projekt i realizacja bazy do *Atlasu gwar wschodniosłowiańskich Białostocczyzny*

Atlas gwar wschodniosłowiańskich Białostocczyzny (1980–2012) to wielotomowa publikacja opracowywana przez zespół z Instytutu Slawistyki PAN przez kilkadziesiąt lat. Pierwszy tom ukazał się w 1980 r., a ostatni – dziesiąty, numerowany jako IV – wyszedł w 2012 r. Materiały do AGWB utrwalają stan gwar wschodniosłowiańskich na północnym wschodzie Polski z lat 50.–70. XX w. Gwar dziś już w znacznym stopniu zintegrowanych z językiem ogólnopolskim, m.in. z miejscowości, które obecnie nie istnieją (z powodu powstania Zalewu Siemianówka). Ta bogata kartoteka jest jedynym w Polsce tak obszernym rejestrem słownictwa i systemu gramatycznego gwar położonych na pograniczu Słowiańszczyzny zachodniej i wschodniej. Z jej zasobów wielokrotnie korzystali naukowcy z Polski, Białorusi i Ukrainy. Jednak ograniczona dostępność kartoteki, która znajduje się w pracowni IS PAN, uniemożliwia rozpowszechnienie tak cennych materiałów.

W 2011 r. złożyłam projekt pt. *Komputerowy system ewidencji archiwaliów do Atlasu gwar wschodniosłowiańskich Białostocczyzny i Słownika bohemizmów*, który został dofinansowany w ramach konkursów Narodowego Programu

Publisher: Institute of Slavic Studies, PAS & The Slavic Foundation
[Wydawca: Instytut Slawistyki PAN & Fundacja Slawistyczna]

Rozwoju Humanistyki (nr projektu 0076/FNiTP/H11/80/201). Jego celem było m.in. stworzenie bazy danych do *Atlasu gwar wschodniosłowiańskich Białostocczyzny* i *Słownika bohemizmów* znajdujących się w zasobach Archiwum Naukowego Instytutu Slawistyki PAN.

Pierwszy etap prac polegał na zeskanowaniu w formacie PDF kartotek do AGWB – prawie 190 000 fiszek. Kolejny etap to stworzenie jednego stanowiska roboczego umożliwiającego przeglądanie i modyfikację bazy danych obsługującej elektroniczną wersję kartoteki do AGWB. Następnie zaplanowałam stworzenie bazy danych i wypełnienie jej materiałami do AGWB.

Wykonywanie skanów i przygotowania do stworzenia bazy odbywały się jednocześnie. Według założeń baza danych ma zawierać nazwy występujące w gwarach wschodniosłowiańskich Białostocczyzny, zapisane w postaci fonetycznej, odpowiadającej pytaniom zapisanym ortograficznie. Docelowo ma ona pomieścić około 250 000 wyrazów zapisanych fonetycznie, dla których odniesieniem są pytania kwestionariusza, będącego podstawą uporządkowania materiału. Pytania są ułożone według działów tematycznych, m.in. I. Części ciała, II. Świat zwierzęcy, III. Budownictwo, VIII. Obróbka włókna, V. Rolnictwo. Np. pod VII 11 ‘gałąź’ znajdziemy formy – *hˈola*, *gˈałęź*, *gałˈenz'a*, *halˈina*, *holˈačka*.

Baza ma umożliwić przeszukiwanie danych według różnych kryteriów:

1. numeru pytania (np. IX 9);
2. numeru wsi (numer punktu);
3. tej samej frazy z różnych pytań kwestionariuszowych (np. *cieląt* z pytania: ‘napój dla cieląt’ i z pytania ‘przywoływanie cieląt’);
4. cząstek wyrazów zapisanych fonetycznie, nie tylko sufiksów i prefiksów (np. *-ova*);
5. wyrazów różniących się tylko położeniem akcentu (np. *pałˈova*, *pałovˈa*);

Dane są porządkowane według:

1. numeru pytania;
2. numeru wsi;
3. form wyrazowych zapisanych fonetycznie w układzie alfabetycznym (*a*, *a̋*, *å*, *b*, *b́*, *c*, *ć*, *c̓*, *č*, *č̓*, *d*, *d́*, *ʒ*, *ʒ́*, *ʒ̓*, *ǯ*, *e*, *ė*, *e̋*, *e̊*, *ě*, *ę*, *f*, *f́* , *g*, *ǵ*, *h*, *h́*, *χ*, *χ́*, *i*, *i͡e*, *i̯*, *k*, *ḱ*, *l*, *ł*, *m*, *ḿ*, *n*, *ŋ*, *n̯*, *ń*, *o*, *ȯ*, *ő*, *o̊*, *o̽*, *o̯*, *o̧*, *p*, *ṕ*, *r*, *ŕ*, *s*, *ś*, *s̓*, *š*, *š̓*, *t*, *t́*, *u*, *ů*, *ü*, *u͡o*, *u̯*, *v*, *v́*, *y*, *y̋*, *y̽*, *y̯*, *y͡e*, *z*, *ź*, *z̓*, *ž*);

4. zakończeń wyrazów.

Początkowo baza miała być dostępna lokalnie z komputera, na którym jest zainstalowana (ewentualnie w sieci przy ustawieniu odpowiednich zabezpieczeń). Dostęp do bazy uzyskiwany byłby za pomocą loginu oraz hasła dla jednego użytkownika.

Autor bazy[1] zdecydował się na zastosowanie systemu do tworzenia baz danych przeznaczonego do ich udostępniania w sieci (alfaISIS, opracowany przez CC Open Computer Systems Ltd. www.cc.com.pl) oraz systemu (WWW-ISIS) przeznaczonego do ich udostępniania w sieci[2].

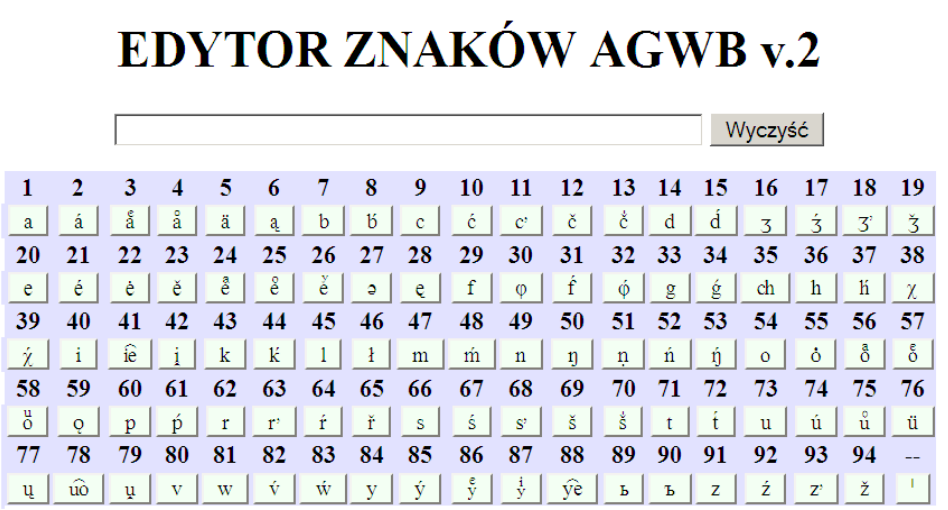
Po przystąpieniu do realizacji koncepcji bazy danych, tak jak przewidywałam, podstawowym problemem okazał się rozszerzony zestaw znaków – UNICODE, który oddawałby wszelkie niuanse fonetyczne. W początkowej fazie korzystano z kroju pisma ZRCola, zaprojektowanego na potrzeby *Ogólnosłowiańskiego atlasu językowego* przez Petera Weissa z Instytutu Języka Słoweńskiego w Lublanie (Inštitut za slovenski jezik Frana Ramovša, Ljubljana). Ten krój jest dostępny, można z niego korzystać w publikacjach. Jednak zestaw znaków ZRColi okazał się niewystarczający. Konieczne były ręczne modyfikacje w zapisie, jak np. tworzenie znaku fonetycznego będącego tak naprawdę kompilacją dwóch lub trzech znaków. Niestety, taki sposób zapisu nazw, oprócz tego, że był wyjątkowo kłopotliwy, uniemożliwiał wdrożenie (zaimplementowanie) w przyszłości w pełni funkcjonalnego modułu wyszukiwania i sortowania form wyrazowych zapisanych w taki sposób. W związku z tym należało ustalić właściwie nowy krój pisma, który nosi nazwę AGWB. Pozwala on na wprowadzanie ustalonych znaków. W trakcie tworzenia bazy wstępny system znaków uległ rozszerzeniu.

Oprócz kroju AGWB stworzony został przy użyciu standardowych narzędzi firmy Microsoft układ klawiaturowy pozwalający na wprowadzanie znaków bezpośrednio z klawiatury wraz z odpowiednią kombinacją klawiszy ALT GR. Oprócz zainstalowanego układu klawiaturowego opracowano edytor w formie strony WWW, w którym użytkownik w specjalnym oknie może wpisywać specjalne znaki, korzystając z ich graficznych modeli.

---

[1] Informacje dotyczące szczegółów technicznych bazy pochodzą od jej autora – dr. Wiesława Glińskiego.

[2] Przykłady zastosowania tego typu rozwiązań można znaleźć m.in. na http://www4.fao.org/faobib/ czy http://biblio.igik.edu.pl/libcat/index.html.



Kolejna istotna dla bazy sprawa to sortowanie danych według porządku znaków fonetycznych. Żadne ze standardowych narzędzi nie mogło poprawnie ułożyć danych zgodnie z ustaloną przeze mnie hierarchią znaków. Ponieważ układałam indeks do wydania papierowego ostatnich tomów AGWB, miałam już doświadczenie w sortowaniu nazw zapisywanych fonetycznie i mogłam podzielić się wiedzą dotyczącą sposobu ustalenia najlepszego wyznaczenia kolejności znaków. Z tego powodu zostały wprowadzone dodatkowe pola, gdzie podano zestaw liczb, które odpowiadają określonemu znakowi, wykorzystywanemu w zapisie fonetycznym. I tak np.

nazwa: *abrˈusek*

ma odpowiedni dla niej zapis cyfrowy określony w bazie jako pole „Dane do sortowania (11): 11177283763053".

Jest jeszcze jeden bardzo ważny dla wyszukiwania danych element, a mianowicie, znak akcentu – tak istotny dla gwar wschodniosłowiańskich, który nie mógł być pominięty we wpisach, co dodatkowo komplikowało sortowanie. Dlatego znak akcentu nie ma reprezentacji liczbowej, aby nie był brany pod uwagę w sortowaniu, ale jednak występował jako znak zapisu.

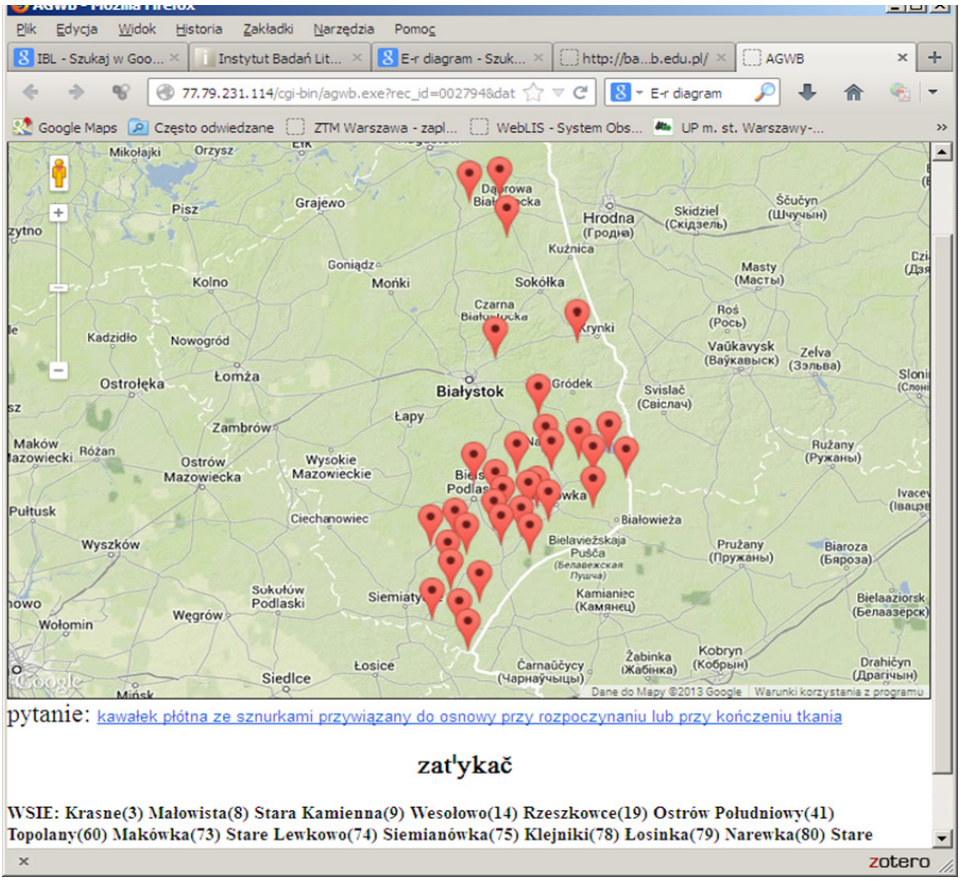
Obecna wersja systemu została wzbogacona o wiele dodatkowych funkcji, m.in. o możliwość pełnego użycia znaków UNICODE, co pozwoliło zastosować w procesie wyszukiwawczym oraz w sortowaniu dowolny układ sztucznych znaków, koniecznych do zapisu fonetycznego. Dla późniejszego procesu wyszukiwawczego ważne jest odnotowanie, że bazy tworzono w systemie, który daje

możliwość wykorzystania odpowiedniego narzędzia indeksującego (Lucene), analogicznego do system wyszukiwawczego Google.

Kiedy został rozstrzygnięty problem fontów niestandardowych, należało zadecydować, jak eksportować dane z arkusza kalkulacyjnego Excel, do którego wpisuję materiał. Każdy rekord był odrębną formą wyrazową. Jeśli pojawiają się formy przypadków zależnych, to wpisuję je do oddzielnego rekordu. Struktura danych zawartych w arkuszu Excel nie ma możliwości wprowadzania danych powtarzalnych, dlatego aby zachować pierwszą postać w bazie, konieczne było powtarzanie w każdym wierszu form wyrazowych dla różnych wartości dotyczących numerów wsi. Takie rozwiązanie odpowiadało strukturze danych zapisanych w fiszkach. Było to jednak nie do przyjęcia w bazie docelowej. W związku z tym należało całkowicie przegrupować dane. W efekcie z kilkunastu tysięcy wierszy w programie Excel, stanowiących pojedyncze rekordy, stworzono w nowym systemie bazę z zaledwie kilkoma tysiącami rekordów. Nowy rekord uwzględnia daną formę wyrazową z odpowiadającym jej oznaczeniem pytania (przez numer pytania i numer działu) oraz z grupą pól powtarzalnych w postaci m.in. numerów wsi, w których te formy występują.

Ostatecznie na system AGWB składają się trzy bazy: baza form wyrazowych (określana jako AGWB), baza zapytań (określana jako PYTANIA) oraz baza wsi (WSIE). W bazie form wyrazowych przechowywane są tylko dane dotyczące samej realizacji fonetycznej nazwy oraz informacje dotyczące numeru działu i numeru pytania. Analogicznie wygląda powiązanie z bazą wsi. Formy wyrazowe są ściśle powiązane z określanymi punktami (numerami wsi), które na fiszkach mają reprezentacje w postaci identyfikatorów cyfrowo-literowych. Baza wsi zawiera poza tym dane pozwalające na określenie położenia geograficznego (długości i szerokości) w standardzie GoogleMaps. Dzięki aplikacji GoogleMaps można w każdej chwili przejść do bazy wsi i odpowiednim formularzu zmodyfikować dane dotyczące bądź to nazwy własnej wsi lub uszczegółowić jej położenie geograficzne (por. ryc. poniżej).

Zastosowanie określonej struktury danych umożliwiło wprowadzenie pól powtarzalnych. Każdy rekord w bazie AGWB ma pole powtarzalne, w którym zawarte są dane dotyczące występowania danej formy wyrazowej ewentualnie form obocznych w różnych wsiach (punktach). Całość traktowana jest jako tzw. grupa powtarzalna i w formularzu wprowadzania danych ma specjalny formularz. Po wypełnieniu danych dotyczących danej fiszki użytkownik może wprowadzić kolejne wystąpienie tego samego pola, ale już dla innej fiszki.



Proces wprowadzania danych dotyczących odwołań do baz pytania i AGWB jest dodatkowo wspomagany przez tzw. tablice haseł wzorcowych. Użytkownik nie musi pamiętać numeru pytania, numeru działu czy też numeru punktu (wsi). W procesie wprowadzania danych w wybranym formularzu ma możliwość podglądu bazy przy danym polu (przycisk LISTA) i wybrania przyciskiem żądanej wartości.

Kontrola poprawności wprowadzanych danych zabezpiecza również użytkownika przed przypadkowym usunięciem rekordu, jeśli wykazuje powiązania z innym rekordem.

Wprowadzanie danych do pól powtarzalnych umożliwia przycisk DODAJ NOWE, który powoduje wyświetlenie całej dodatkowej grupy pól do wypełnienia.

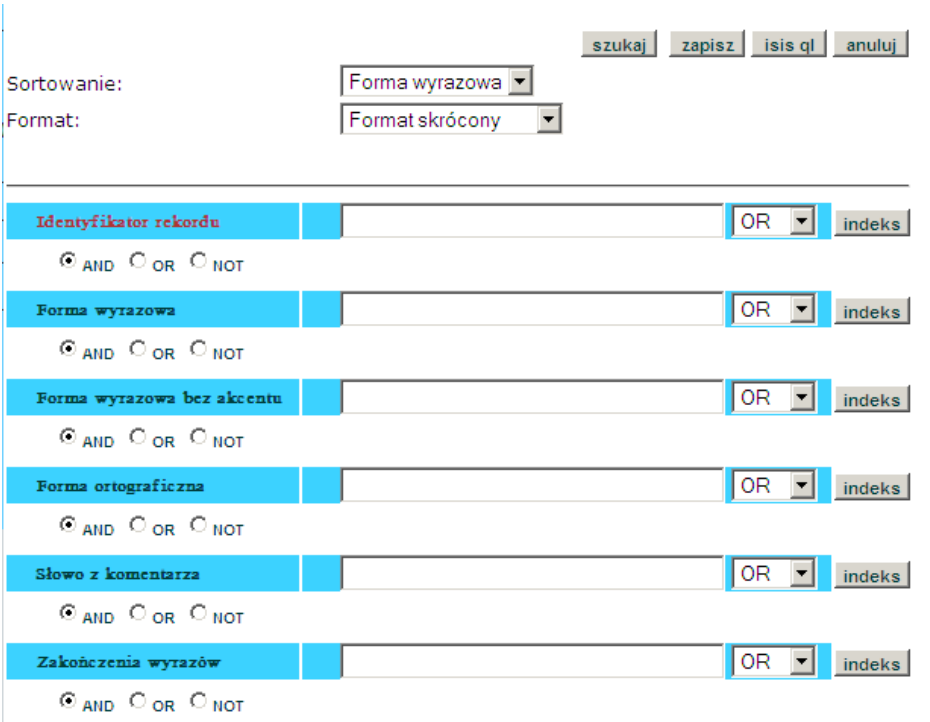
Kolejnym krokiem zwiększającym funkcjonalność bazy była możliwość wprowadzenia wyświetlania skanów fiszek. Na wprowadzenie do bazy skanów fiszek zdecydowałam się w późniejszym etapie projektowania, dlatego na razie niewiele rekordów w bazie AGWB jest z nimi połączonych.

Dla użytkownika interesujący jest mechanizm docierania do danych. Zaprojektowano dwa sposoby przeszukiwania (dwa moduły komunikacji systemu z użytkownikiem). Pierwszy (uproszczony) uwzględnia jedno pole, drugi zaś (zaawansowany) zbudowano z wielu pól wyszukiwawczych. Pole wyszukiwawcze w wyszukiwaniu uproszczonym uwzględnia m.in. słowa kluczowe z całego rekordu: formę wyrazową (zapisaną w kroju AGWB), słowa kluczowe z pytania, numer działu pytania i wsi, komentarze itd. W przypadku zaawansowanego sposobu wyszukiwawczego możliwe jest wyszukiwanie według następujących kryteriów:

- identyfikatora rekordu,
- formy wyrazowej,
- formy wyrazowej bez akcentu,
- formy ortograficznej,
- słowa z komentarza,
- zakończeń wyrazów,
- numeru wsi,
- numeru działu,
- numeru pytania,
- nazwy wsi,
- słowa kluczowego z pytania,
- słowa kluczowego z rekordu.

Pola wyszukiwawcze mogą być połączone operatorami AND, OR lub NOT (domyślnie są to operatory AND). Wewnątrz danego pola terminy mogą być łączone operatorem OR lub AND (domyślnie OR).

Do wymienionych funkcji bazy należy dodać możliwość wyświetlania danych oraz ich zapisywania w postaci dowolnych typów plików tekstowych. Aby uniknąć przeciążenia systemu, zaleca się ustalenie maksymalnej liczby znalezionych rekordów na poziomie 500–800. Wynik procesu wyszukiwania jest wzbogacony dodatkowo możliwością zapisania wyników w edytorze tekstu (Word), arkuszu kalkulacyjnym (Excel) lub jako pliku dla dowolnej przeglądarki www (plik html). Ich wybór zależy od potrzeb i preferencji końcowego użytkownika i może być z powodzeniem wykorzystywany w przygotowaniu wersji papierowej słownika.



Obecnie serwer z uzupełnianą bazą AGWB znajduje się w Instytucie Slawistyki PAN. Nie została jeszcze podjęta ostateczna decyzja co do dostępności bazy na etapie wprowadzania danych. Najprawdopodobniej dostęp będzie limitowany aż do ukończenia całości. Wykupiony serwer online uniezależnia od serwera stacjonarnego, co umożliwia prowadzenie prac w różnych miejscach.

Całkowite ukończenie bazy ma nastąpić w styczniu 2015 r. Pozostało do wpisania ok. 2/3 materiału. Wpisywanie jest benedyktyńską pracą, która jednak przyniesie oczekiwany efekt, opracowywana baza bez wątpienia będzie bowiem dużym udogodnieniem dla wszystkich, którzy chcą wykorzystywać materiał do AGWB, mający obecnie charakter historyczny i unikatowy.

Bibliografia

Glinka, S., Obrębska-Jabłońska, A., Siatkowski, J., & Maryniakowa, I. (Red.) (1980–2012). *Atlas gwar wschodniosłowiańskich Białostocczyzny* (T. 1–10). Wrocław: Zakład Narodowy im. Ossolińskich; Warszawa: Slawistyczny Ośrodek Wydawniczy.

Project and realization of database for the *Atlas of East Slavic dialects of Bialystok Region*

Summary

The article presents issues concerning the project of creation of the *Atlas of East Slavic dialects of Bialystok Region* (*Atlas gwar wschodniosłowiańskich Białostocczyzny* – AGWB) and its realization. The crucial issues are shown from linguistic, information and computer points of view, taking into account the special features of dialectal material. Information referring to the *Atlas of East Slavic dialects of Bialystok Region*, a publication significant for Slavic linguistic geography, is also presented.

The consecutive stages of preparing the base, i.a. preliminary choice of computer tools, suggested database structure, its subsequent modifications along with the E-R model and initial stages of system implementation (export of data, the choice of database administration system), are presented in the text. The problem of creation/modification of the phonetic alphabet (so-called ZRCola typeface vs. AGWB typeface) and the aspect of sorting data saved in this alphabet constitute separate issues for consideration.

Modifications of data representation system in relation to the initial objectives are discussed in the article, taking into account the future needs of users (connection with GoogleMaps application, hipermedial connections with the system of scanned source materials, etc.).

Keywords: dialectal atlas; data base; digitalization of archives

Słowa kluczowe: atlas gwarowy; baza danych; digitalizacja archiwaliów