Zbigniew Handzel[1]
Mirosław Gajer[2]

W kierunku automatycznej klasyfikacji języków naturalnych

Towards an automatic classification of natural languages

Streszczenie: Klasyfikacja języków naturalnych jest jednym z głównych zadań językoznawstwa. Spośród różnych typów klasyfikacji języków najbardziej wiarygodną i miarodajną wydaje się być klasyfikacja typologiczna, która łączy języki w jednostki większego rzędu na podstawie podobieństwa ich cech strukturalnych. Podobieństwo typologiczne języków może być wynikiem zarówno ich pochodzenia od wspólnego przodka, czyli prajęzyka, jak i występujących zapożyczeń międzyjęzykowych dotyczących zarówno leksyki, jak i struktur składniowych. W artykule zamieszczono propozycję budowy systemu przeznaczonego do realizacji automatycznej klasyfikacji języków naturalnych ze względu na ich stopień podobieństwa typologicznego.
Opracowany przez autorów system uwzględnia obecnie 72 języki należące głównie do indoeuropejskiej rodziny językowej. W systemie uwzględniono ponadto kilka języków należących do innych rodzin językowych oraz wybrane języki sztuczne typu naturalistycznego. Autorzy zaprezentowali program komputerowy służący do wyznaczania liczbowej miary stopnia wzajemnego podobieństwa systemów zaimków osobowych występujących w różnych językach świata. W przyszłości planowana jest budowa analogicznych systemów przeznaczonych do wyznaczania miary podobieństwa języków na podstawie automatycznej analizy wzorców koniugacyjnych czasowników oraz wzorców deklinacyjnych rzeczowników i przymiotników wybranych języków.

[1] Dr inż. Prof. Wyższej Szkoły Ekonomii i Informatyki w Krakowie.
[2] Dr inż., AGH Akademia Górniczo-Hutnicza, Wydział EAlilB, Katedra Informatyki Stosowanej.

Słowa kluczowe: lingwistyka komputerowa, przetwarzanie języka naturalnego, klasyfikacja języków.

Summary: Classification of natural languages is one of the main tasks of linguistics. Of the various types of language classification, the most reliable and authoritative seems to be the typological classification, which combines languages into units of a higher order on the basis of similarity of their structural features. The typological similarity of languages may be a result of both their origin from a common ancestor, i.e. a proto-language, and interlingual borrowings concerning both lexis and syntactic structures. The paper presents a proposal for the construction of a system intended for the automatic classification of natural languages according to their degree of typological similarity.
The system developed by the authors currently includes 72 languages belonging mainly to the Indo-European language family. The system also includes several languages belonging to other language families and selected artificial languages of a naturalistic type. The authors have presented a computer programme for determining a numerical measure of the degree of mutual similarity between the systems of personal pronouns occurring in different languages of the world. In the future it is planned to build analogous systems to determine the measure of similarity between languages on the basis of automatic analysis of verb conjugation patterns and declension patterns of nouns and adjectives of selected languages.

Keywords: computational linguistics, natural language processing, language classification.

*JEL classification codes***:** *C6, C60.*

Wprowadzenie

Jednym z podstawowych zadań stojących przed dyscypliną językoznawstwa jest klasyfikowanie materiału językowego istniejącego na kuli ziemskiej zarówno obecnie, jak i w znanej nam z badań historycznych odległej przeszłości[3]. Wyróżnia się przy tym trzy odmienne systemy klasyfikacji języków, do których należą: klasyfikacja genetyczna, klasyfikacja geograficzna oraz klasyfikacja typologiczna.

[3] Majewicz A. F., *Języki świata i ich klasyfikowanie*, Państwowe Wydawnictwo Naukowe, Warszawa, 1989.

System genetycznej klasyfikacji języków jest historycznie najstarszy, a jego podstawę stanowi domniemanie pochodzenia danej grupy językowej od wspólnego im wszystkim przodka – prajęzyka. Na przykład takim wspólnym przodkiem dla wszystkich języków romańskich był używany w starożytności na znacznych obszarach Europy język łaciński[4]. Podobnie, uważa się, że wszystkie obecnie istniejące, a także i wymarłe w przeszłości (na przykład język staro-cerkiewno-słowiański lub język połabski), języki słowiańskie wywodzą się od wspólnego przodka, którym był rekonstruowany przez badaczy język prasłowiański[5]. Niestety, język ten nie posiadał swej formy pisanej, zatem żaden jego zapis z oczywistych powodów nie dotarł do naszych czasów, a obecna wiedza o nim opiera się wyłącznie na dokonanych przez lingwistów domniemanych rekonstrukcjach.

Często klasyfikacja genetyczna języków porównywana jest do klasyfikacji organizmów żywych, które łączone są w odpowiednie jednostki systematyczne na podstawie przewidywanego ich pochodzenia od wspólnego przodka. Taki tok rozumowania nie jest jednak do końca słuszny ponieważ w przypadku języków naturalnych przepływ informacji ma charakter nie tylko „pionowy", (czyli w kierunku od przodków do potomków), ale równie często występuje przekaz „poziomy" w postaci zapożyczeń (najczęściej w sferze leksyki) od sąsiednich języków, które w ogólnym przypadku wcale od rozważanego wspólnego przodka nie muszą bynajmniej pochodzić[6].

W wyniku tego rodzaju zapożyczeń języki należące nawet do zupełnie odrębnych rodzin językowych mogą się do siebie wzajemnie nawet w bardzo dużym stopniu upodobnić. Dawniej w językoznawstwie mówiono powszechnie o chińsko-tybetańskiej rodzinie językowej, w której wyróżniano cztery grupy językowe: chińską, tybetańską, birmańską i tajską[7]. Obecnie, według najnowszych poglądów, mamy w istocie do czynienia z czterema różnymi rodzinami językowymi, a należące do nich języki upodobniły się w znacznym stopniu do siebie w wyniku wzajemnych kontaktów ich użytkowników (są to wszystko języki w bardzo wysokim stopniu analityczne, monosylabiczne i toniczne). Taki stan rzeczy powoduje, że jest rzeczą sensowną podjąć próbę klasyfikowania języków pod względem

4) Mańczak W., *Języki romańskie* [w:] *Języki indoeuropejskie* (pod redakcją Leszka Bednarczuka), tom II, Państwowe Wydawnictwo Naukowe, Warszawa, 1988.

5) Majewicz A. F., *Języki świata i ich klasyfikowanie*, Państwowe Wydawnictwo Naukowe, Warszawa, 1989.

6) Matisoff J. A., *Zagrożona różnorodność: języki i formy życia*, Świat Nauki, nr 10, 2002, ss. 66-73.

7) Künstler J. M., *Języki chińskie*, Warszawa, Wydawnictwo Akademickie DIALOG, 2000.

analizy obszarów ich występowania, co dokonywane jest w ramach klasyfikacji geograficznej języków, poprzez łączenie ich w tzw. ligi językowe.

Jednak najbardziej miarodajnym typem klasyfikacji języków wydaje się być klasyfikacja typologiczna, która uwzględnia podobieństwa pomiędzy językami, niezależnie od tego, czy są one wynikiem pochodzenia rozpatrywanych języków od wspólnego im przodka – prajęzyka, czy też powstały w wyniku geograficznej bliskości ich występowania.

Automatyzacja klasyfikacji języków

Według obiegowych opinii język czeski jest bardzo podobny do języka słowackiego i zapewne jest w tym stwierdzeniu sporo racji, ponieważ użytkownicy obu wymienionych języków nie mają zwykle najmniejszych trudności ze wzajemną komunikacją językową. Z kolei język polski jest do języka czeskiego również w wysokim stopniu podobny, ale zapewne już nie tak bardzo jak język słowacki. Należy przy tym zauważyć, że wszystkie trzy wymienione języki należą do jednej podgrupy języków zachodniosłowiańskich. Natomiast język chorwacki, który należy do grupy języków południowosłowiańskich, jest do rozważanego tutaj języka czeskiego podobny już w mniejszym stopniu. W jeszcze mniejszym stopniu język czeski jest podobny do języka rosyjskiego, który jest największym pod względem liczby użytkowników językiem należącym do wschodniosłowiańskiej grupy językowej.

Są to wszystko kwestie dość oczywiste, które wyczuwane są przez użytkowników języków słowiańskich w sposób intuicyjny. Jednak sporym wyzwaniem jest próba ujęcia tych intuicyjnie wychwytywanych zależności w swego rodzaju liczbowe miary stopnia wzajemnego podobieństwa języków. Gdyby tego rodzaju precyzyjną miarę liczbową udało się utworzyć, wówczas klasyfikację języków można byłoby przeprowadzić w wielu wypadkach w sposób o wiele bardziej wiarygodny. Być może tego rodzaju liczbowa miara stopnia wzajemnego podobieństwa języków mogłaby być ważnym argumentem pozwalającym w pewnych przypadkach w definitywny sposób rozstrzygnąć, czy mamy do czynienia z całkowicie odrębnym i niezależnym językiem, czy jedynie z jakąś formą dialektalną innego języka.

W rozważanym kontekście wystarczy wspomnieć tylko toczone niegdyś zawzięcie spory o to, czy kaszubski jest odrębnym językiem, czy jedynie jednym z wielu dialektów języka polskiego. Jeśli uświadomimy sobie fakt, że dla przeciętnego użytkownika języka polskiego bliski mu język słowacki jest znacznie lepiej zrozumiały niż kaszubski, wówczas kwestia klasyfikacji ka-

szubskiego jako odrębnego języka wydaje się raczej oczywista (nikt przecież nie ma wątpliwości, że słowacki jest językiem odrębnym od polskiego).

Tego rodzaju sytuacji znanych jest w językoznawstwie dosłownie bez liku i niestety w większości przypadków za ostatecznym rozwiązaniem przemawiają kwestie natury pozajęzykowej – głównie politycznej. Przykładowo, na należących do Japonii wyspach Riukiu używany jest lokalny język, określany mianem języka ryukyu, który w Japonii uważany jest po prostu za dialekt języka japońskiego. Tymczasem badania naukowe stanowisko takie zdecydowanie wykluczają, ponieważ jakiekolwiek (nawet bardzo odległe) genetyczne pokrewieństwo języka japońskiego z językiem ryukyu w ogóle nie zostało udowodnione, a ponadto oba wymienione języki są wzajemnie dla ich użytkowników całkowicie niezrozumiałe[8].

W świetle powyższych uwag prowadzenie badań mających na celu utworzenie systemu dającego precyzyjną miarę stopnia podobieństwa języków wydaje się być dostatecznie mocno uzasadnione. Podstawę opracowywanego przez autorów systemu stanowi typologiczna klasyfikacja języków, która może być przeprowadzana na wielu równoległych płaszczyznach: fonologicznej, syntaktycznej i semantycznej[9]. W pierwszym etapie prowadzonych badań autorzy skoncentrowali się na klasyfikacji typologicznej systemów zaimków osobowych języków naturalnych.

Klasyfikacja systemów zaimków osobowych

Zaimki osobowe w językoznawstwie zaliczane są do tzw. uniwersaliów językowych, to znaczy stanowią kategorię gramatyczną, która musi być obecna w każdym języku naturalnym. Jak dotychczas nie jest znany żaden język naturalny, który by jakiegoś, nawet bardzo uproszczonego systemu zaimków osobowych w ogóle nie posiadał. W wielu różnych językach świata zaimki osobowe odmieniają się przez osoby gramatyczne, a także niekiedy przez liczby i rodzaje gramatyczne, co sprawia, że systemy zaimków osobowych poszczególnych języków naturalnych mogą bardzo różnić się od siebie[10].

---

8) Majewicz A. F., *Języki świata i ich klasyfikowanie*, Państwowe Wydawnictwo Naukowe, Warszawa 1989.

9) Milewski T., *Językoznawstwo*, Wydawnictwo Naukowe PWN, Warszawa 2004.

10) Arnold D., Balkan L., Meijer S., Humphreys R. L., Sadler L., *Machine translation: an introductory guide, NCC Blackwell*, London 1994.

W ogólnym przypadku system zaimków osobowych dla dowolnego języka naturalnego, w którym mogą maksymalnie wystąpić trzy liczby gramatyczne, może zostać przedstawiony za pomocą następującej tabeli.

Tabela 1. Ogólny schemat systemu zaimków osobowych języka naturalnego

	Liczba pojedyncza	**Liczba podwójna**	**Liczba mnoga**
Osoba pierwsza	ja (mężczyzna)	my dwaj (mężczyźni)	my (mężczyźni)
	ja (kobieta)	my dwie (kobiety)	my (kobiety)
Osoba druga	ty (mężczyzna)	wy dwaj (mężczyźni)	wy (mężczyźni)
	ty (kobieta)	wy dwie (kobiety)	wy (kobiety)
Osoba trzecia	on (mężczyzna)	oni dwaj (mężczyźni)	oni (mężczyźni)
	ona (kobieta)	one dwie (kobiety)	one (kobiety)
	ono (dziecko)	ich dwoje (dzieci)	one (dzieci)

Źródło: opracowanie własne.

Jak wynika z tab. 1, system zaimków osobowych języka naturalnego, w którym występują nie więcej niż trzy liczby gramatyczne, może maksymalnie liczyć aż 21 różnych form wyrazowych. Jednak język, który posiadałby tak wielką mnogość odmiennych form wyrazowych dla zaimków osobowych prawdopodobnie nigdzie nie istnieje i nigdy też w przeszłości nie istniał, ponieważ tak rozbudowany system zaimków byłby bardzo nieekonomiczny i obciążałby zbytnio pamięć osób posługujących się danym językiem. W związku z powyższym w przypadku konkretnych języków mamy zwykle do czynienia ze znacznym uproszczeniem przedstawionego w tab. 1 ogólnego schematu.

Przykładowo, skrajnie prosty system zaimków osobowych występuje w języku perskim, zaliczanym do indoeuropejskiej rodziny językowej. System zaimków osobowych języka perskiego został przedstawiony w tab. 2. Rozważany system zaimków osobowych jest tak prosty ponieważ we współczesnym języku perskim, (w wyniku jego historycznego rozwoju), całkowicie zanikła kategoria rodzaju gramatycznego.

Tabela 2. Zaimki osobowe współczesnego języka perskiego (uproszczona transkrypcja fonetyczna)

	Liczba pojedyncza	Liczba podwójna	Liczba mnoga
Osoba pierwsza	man	-	ma
	-	-	-
Osoba druga	to	-	szoma
	-	-	-
Osoba trzecia	o	-	anha
	-	-	-
	-	-	-

Źródło: opracowanie własne.

Z kolei w tab. 3 przedstawiono wygląd systemu zaimków osobowych w klasycznym języku arabskim (zastosowano uproszczoną transkrypcję fonetyczną). W rozważanym wypadku wielość form wyrazowych zaimków osobowych klasycznego języka arabskiego wynika z występowania w tym języku kategorii gramatycznej liczby podwójnej (tzw. dualisu). Dodatkowo, w wielu przypadkach występują odrębne formy zaimków (w drugiej osobie gramatycznej) dla rodzaju męskiego i żeńskiego (kategoria gramatyczna rodzaju nijakiego nie jest w klasycznym języku arabskim, podobnie jak i w innych językach semickich, w ogóle znana)[11].

Tabela 3. Zaimki osobowe klasycznego języka arabskiego (uproszczona transkrypcja fonetyczna)

	Liczba pojedyncza	Liczba podwójna	Liczba mnoga
Osoba pierwsza	ana	nahnu	nahnu
	ana	nahnu	nahnu

[11] Danecki J., *Współczesny język arabski i jego dialekty*, Wydawnictwo Akademickie DIALOG, Warszawa 2000.

Osoba druga	anta	antuma	antum
	anti	antuma	antunna
Osoba trzecia	hua	huma	hum
	hija	huma	hunna
	-	-	-

Źródło: opracowanie własne.

W celu budowy systemu pomiaru odległości pomiędzy systemami zaimków osobowych różnych języków należy najpierw rozwiązać problem przekształcenia danych lingwistycznych w dane o charakterze numerycznym.

Przekształcenie danych lingwistycznych do postaci numerycznej

Autorzy niniejszego artykułu zastosowali następujący sposób konwersji danych o charakterze lingwistycznym do postaci numerycznej. W tym celu analizowane są kolejne formy wyrazowe zaimków osobowych danego języka w następującym porządku:

1. ja (mężczyzna)
2. ja (kobieta)
3. ty (mężczyzna)
4. ty (kobieta)
5. on (mężczyzna)
6. ona (kobieta)
7. ono (dziecko)
8. my dwaj (mężczyźni)
9. my dwie (kobiety)
10. wy dwaj (mężczyźni)
11. wy dwie (kobiety)
12. oni dwaj (mężczyźni)
13. one dwie (kobiety)
14. ich dwoje (dzieci)
15. my (mężczyźni)
16. my (kobiety)
17. wy (mężczyźni)

18. wy (kobiety)
19. oni (mężczyźni)
20. one (kobiety)
21. one (dzieci)

Z powyższych danych lingwistycznych tworzony jest wektor o 21 składowych. Każda składowa rozważanego wektora może przyjąć wartość z następującego zbioru wartości numerycznych {-1, 0, 1}. Jeśli dana forma wyrazowa występuje po raz pierwszy, wówczas na skojarzonej z nią pozycji wektora wartości numerycznych pojawia się jedynka. W przypadku przeciwnym, czyli jeśli dana forma wyrazowa wystąpiła już uprzednio, wówczas odpowiednia składowa wektora wartości numerycznych przyjmuje wartość zero. Natomiast jeśli dana forma zaimka w ogóle nie występuje, wówczas na skojarzonej z nią pozycji wektora pojawi się minus jedynka.

W przypadku języka polskiego na poszczególnych pozycjach pojawią się następujące formy wyrazowe zaimków osobowych:

1. ja
2. ja
3. ty
4. ty
5. on
6. ona
7. ono
8. (brak formy wyrazowej)
9. (brak formy wyrazowej)
10. (brak formy wyrazowej)
11. (brak formy wyrazowej)
12. (brak formy wyrazowej)
13. (brak formy wyrazowej)
14. (brak formy wyrazowej)
15. my
16. my
17. wy
18. wy
19. oni
20. one
21. one

W związku z powyższym dla systemu zaimków osobowych języka polskiego wygenerowany zostanie następujący wektor o 21 składowych:

[1, 0, 1, 0, 1, 1, 1, -1, -1, -1, -1, -1, -1, -1, 1, 0, 1, 0, 1, 1, 0]

Rozważany wektor jest następnie poddawany normalizacji, tak aby jego długość była jednostkowa. Omawiana operacja polega na podzieleniu każdej ze składowych rozpatrywanego wektora przez jego długość, czyli przez pierwiastek z sumy kwadratów poszczególnych współrzędnych.

Odległość pomiędzy wektorami wyznaczonymi dla systemów zaimków osobowych dwóch różnych języków definiujemy jako ich iloczyn skalarny. Otrzymujemy wówczas liczbę z przedziału [-1, 1]. Jeżeli teraz do uzyskanego wyniku dodamy jeden, uzyskamy liczbę z przedziału [0,2]. Teraz wystarczy uzyskany wynik podzielić przez dwa, aby miara odległości pomiędzy systemami zaimków osobowych badanych języków była liczbą z przedziału [0, 1]. Uzyskanie wartości równej jeden oznacza maksymalne podobieństwo systemów zaimków osobowych badanych języków. Z kolei uzyskanie wartości zero oznacza brak jakiegokolwiek ich podobieństwa (wartość raczej niemożliwa do uzyskania w praktyce).

Implementacja systemu w języku Python

Program wyznaczający miarę stopnia podobieństwa systemów zaimków osobowych w dwóch różnych językach został napisany przez autorów w języku Python. Celowo wybrany został nowoczesny język programowania, charakteryzujący się przejrzystą i prostą składnią, podlegający procesowi interpretacji [8]. Język ten wykorzystywany jest obecnie powszechnie w dziedzinie sztucznej inteligencji, uczenia maszynowego i przetwarzania języka naturalnego[12].

Opracowany przez autorów w języku Python program został wyposażony w graficzny interfejs użytkownika, który został utworzony z wykorzystaniem biblioteki „tkinter". Na rys. 1 zamieszczono zrzut ekranu przedstawiający widok okna głównego programu.

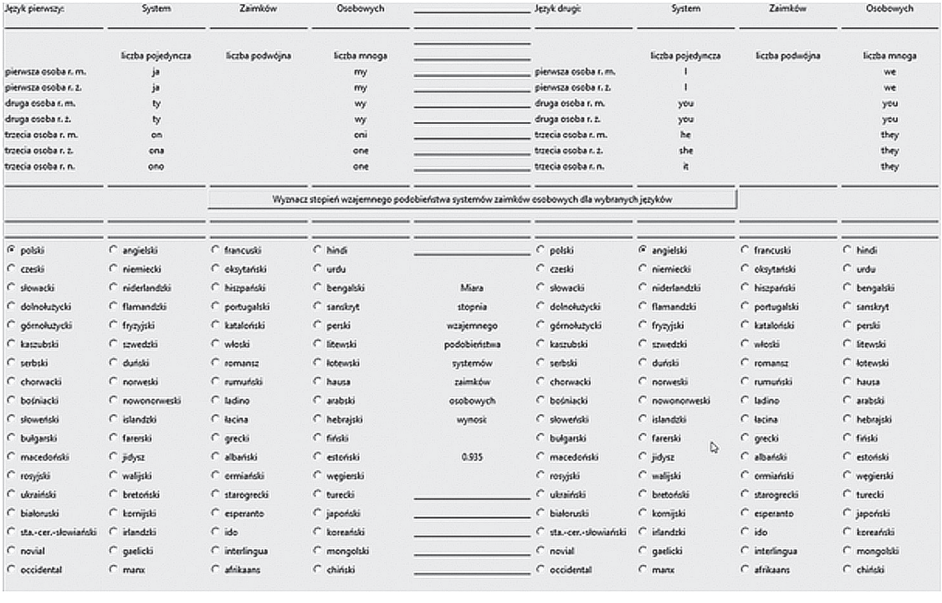
Za pomocą przycisków opcji wyboru użytkownik dokonuje wyboru języka, dla którego mają zostać wyświetlone istniejące w nim zaimki osobowe. Na chwilę obecną w opracowanym przez autorów systemie uwzględniono 72 języki, należące głównie do indoeuropejskiej rodziny językowej. Spośród języków słowiańskich uwzględniono języki takie jak: polski, czeski, słowacki, dolnołużycki, górnołużycki, kaszubski, serbski, chorwacki, bośniacki, słoweński, bułgarski, macedoński, rosyjski, ukraiński, białoruski i staro-cerkiewno-słowiański. Z kolei z grupy języków germańskich uwzględniono języki takie jak: angielski, niemiecki, niderlandzki, flamandzki, fryzyjski, szwedzki, duń-

[12] Raschka S., *Python – uczenie maszynowe*, Wydawnictwo HELION, Gliwice, 2018.

ski, norweski, nowonorweski (nynorsk), islandzki, farerski, afrikaans i jidysz. Podobnie z grupy języków romańskich w systemie ujęto takie języki jak: francuski, oksytoński, hiszpański, portugalski, kataloński, włoski, romansz, rumuński, ladino i język łaciński.

Utworzony przez autorów system obejmuje także inne języki indoeuropejskie, nie należące do uprzednio wymienionych grup językowych (słowiańskiej, germańskiej i romańskiej). Są to następujące języki: walijski, bretoński, kornijski, irlandzki, gaelicki, manx, grecki, starogrecki, albański, ormiański, hindi, urdu, bengalski, sanskryt, litewski, łotewski i perski. W rozważanym systemie uwzględniono także wybrane języki sztuczne typu naturalistycznego (wzorowane na języku łacińskim i innych współczesnych językach Europy), takie jak esperanto, ido, novial, occidental i interlingua. Ponadto system uwzględnia kilka języków świata nie należących do indoeuropejskiej rodziny językowej. Są to następujące języki: arabski, hebrajski, fiński, estoński, węgierski, turecki, japoński, koreański, mongolski i chiński.

Rysunek 1. Widok okna programu wyznaczającego stopień podobieństwa systemów zaimków osobowych wybranych języków świata



Źródło: opracowanie własne.

Współpraca użytkownika z rozważanym programem polega na dokonaniu wyboru pierwszego języka, poprzez kliknięcie odpowiedniego przycisku wyboru opcji. Na rys. 2 zamieszczono zrzut ekranu w przypadku, gdy użytkownik jako język pierwszy wybrał język górnołużycki.

Rysunek 2. Selekcja pierwszego języka poprzez odpowiedni przycisk wyboru opcji

Źródło: opracowanie własne.

Po dokonaniu wyboru pierwszego języka w głównym oknie programu (u góry po lewej stronie) wyświetlone zostają wszystkie formy wyrazowe zaimków osobowych występujących w danym języku, co pokazano na rys. 3.

Rysunek 3. System zaimków osobowych wybranego przez użytkownika języka górnołużyckiego

Język pierwszy:	System	Zaimków	Osobowych
	liczba pojedyncza	liczba podwójna	liczba mnoga
pierwsza osoba r. m.	ja	mój	my
pierwsza osoba r. ż.	ja	mój	my
druga osoba r. m.	ty	wój	wy
druga osoba r. ż.	ty	wój	wy
trzecia osoba r. m.	wón	wonaj	woni
trzecia osoba r. ż.	wona	wonej	wone
trzecia osoba r. n.	wono	wonej	wone

Źródło: opracowanie własne.

W etapie kolejnym użytkownik poprzez kliknięcie odpowiedniego przycisku wyboru opcji dokonuje wyboru drugiego języka. Na rys. 4 pokazano przypadek, gdy użytkownik jako drugi język wybrał współczesny język hebrajski.

Rysunek 4. Selekcja drugiego języka poprzez odpowiedni przycisk wyboru opcji

Źródło: opracowanie własne.

Po dokonaniu wyboru drugiego języka w głównym oknie programu (u góry po prawej stronie) wyświetlone zostają wszystkie formy wyrazowe zaimków osobowych występujących w danym języku, co pokazano na rys. 5.

Rysunek 5. System zaimków osobowych wybranego przez użytkownika języka hebrajskiego

	liczba pojedyncza	liczba podwójna	liczba mnoga
pierwsza osoba r. m.	ani		anachnu
pierwsza osoba r. ż.	ani		anachnu
druga osoba r. m.	ata		atem
druga osoba r. ż.	at		aten
trzecia osoba r. m.	hem		hem
trzecia osoba r. ż.	hen		hen
trzecia osoba r. n.			

Źródło: opracowanie własne.

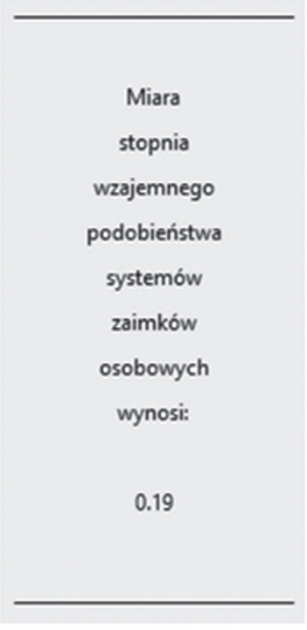
W kolejnym etapie użytkownik musi kliknąć przycisk z napisem „Wyznacz stopień wzajemnego podobieństwa systemów zaimków osobowych dla wybranych języków", co pokazano na rys. 6

Rysunek 6. Przycisk uruchamiający funkcje wyznaczającą stopień podobieństwa systemów zaimków osobowych wybranych uprzednio języków

Źródło: opracowanie własne.

Po kliknięciu w rozważany przycisk wyświetlany jest automatycznie wyliczony rezultat stopnia wzajemnego podobieństwa systemów zaimków osobowych dla wybranych przez użytkownika języków, co zobrazowano na rys. 7.

Rysunek 7. Uzyskany rezultat miary stopnia podobieństwa systemów zaimków osobowych języków górnołużyckiego i hebrajskiego



Źródło: opracowanie własne.

W rozpatrywanym powyżej przykładzie stopień wzajemnego podobieństwa systemów zaimków osobowych języków górnołużyckiego i hebrajskiego jest stosunkowo niewielki, ponieważ wynosi zaledwie 0,19 (przy maksymalnej możliwej wartości równej jeden). Jednak uzyskany rezultat nie powinien być żadnym zaskoczeniem, gdyż – po pierwsze – mamy tutaj do czynienia z dwoma bardzo odległymi od siebie językami (słowiańskim i semickim), a po wtóre, systemy zaimków osobowych rozważanych języków różnią się pomiędzy sobą w istotny sposób. W przypadku języka górnołużyckiego występuje kategoria gramatyczna liczby podwójnej, która we współczesnym języku hebrajskim już w zasadzie całkowicie zanikła (poza nielicznymi przypadkami rzeczowników oznaczających na przykład parzyste części ciała). Z kolei w przypadku języka hebrajskiego odmienne formy wyrazowe zaimków osobowych pojawiają się w drugiej osobie rodzaju żeńskiego zarówno w liczbie pojedynczej, jak i mnogiej.

Zakończenie

Opisany powyżej mechanizm automatycznego wyznaczania stopnia wzajemnego podobieństwa systemów zaimków osobowych różnych języków stanowi pierwszy etap budowy większego systemu służącego do przeprowadzania automatycznej klasyfikacji typologicznej języków naturalnych. Oczywiście analiza systemów zaimków osobowych języków stanowi tylko jeden z przyczynków do ich klasyfikacji typologicznej, która aby mogła być uznana za pełną, musi uwzględniać jeszcze wiele innego rodzaju cech badanych języków naturalnych.

W związku z powyższym w dalszym etapie badań autorzy planują budowę analogicznych systemów, które mierzyłyby odległości pomiędzy językami naturalnymi na podstawie analizy systemów koniugacyjnych występujących w nich czasowników oraz systemów deklinacyjnych rzeczowników i przymiotników. Autorzy zamierzają w tym wypadku oprzeć się na powszechnie uznawanej w językoznawstwie tzw. liście Swadesha, obejmującej wykaz wyrazów najczęściej używanych w językach naturalnych. W rozważanym wypadku dla czasowników znajdujących się na liście Swadesha porównywane byłyby ich wzorce koniugacyjne, na podstawie czego generowana byłaby analogiczna miara liczbowa, jak miało to miejsce w przypadku opisanej w niniejszym artykule analizy systemów zaimków osobowych. Z kolei w przypadku języków naturalnych, w których występuje odmiana przez przypadki, porównywane byłyby w analogiczny sposób wzorce deklinacyjne rzeczowników i przymiotników zawartych na liście Swadesha.

Literatura

1. Arnold D., Balkan L., Meijer S., Humphreys R. L., Sadler L., *Machine translation: an introductory guide, NCC Blackwell*, London, 1994.
2. Dalewska-Greń H., *Języki słowiańskie*, Wydawnictwa Naukowe PWN, Warszawa, 2002.
3. Danecki J., *Współczesny język arabski i jego dialekty*, Wydawnictwo Akademickie DIALOG, Warszawa, 2000.
4. Künstler J. M., *Języki chińskie*, Warszawa, Wydawnictwo Akademickie DIALOG, 2000.
5. Majewicz A. F., *Języki świata i ich klasyfikowanie*, Państwowe Wydawnictwo Naukowe, Warszawa, 1989.
6. Mańczak W., *Języki romańskie* [w:] „Języki indoeuropejskie"(pod redakcją Leszka Bednarczuka), tom II, Państwowe Wydawnictwo Naukowe, Warszawa, 1988.

7. Matisoff J. A., *Zagrożona różnorodność: języki i formy życia*, Świat Nauki, nr 10, 2002, ss. 66-73.
8. Miles R., *Python – zacznij programować*, Wydawnictwo HELION, Gliwice, 2019.
9. Milewski T., *Językoznawstwo*, Wydawnictwo Naukowe PWN, Warszawa, 2004.
10. Raschka S., *Python – uczenie maszynowe*, Wydawnictwo HELION, Gliwice, 2018.