Lětopis 72 2025

Zeitschrift für Sorabistik und vergleichende Minderheitenforschung
Časopis za sorabistiku a přirunowace mjeńšinowe slědźenje
Casopis za sorabistiku a psirownujuce mjeńšynowe slěźenje
Journal for Sorbian and Comparative Minority Studies

Bartłomiej Szawulak, Tadeusz Lewaszkiewicz, Piotr Formanowicz

# Operationen an Wortmengen als Mittel zur Bestimmung des Ähnlichkeitsgrades zwischen verwandten Sprachen (am Beispiel des Niedersorbischen und des Polnischen)

Die slawischen Sprachen zeichnen sich im Vergleich mit anderen indoeuropäischen Sprachen durch einen hohen Ähnlichkeitsgrad aus, der z. B. im gemeinsamen Wortschatz ersichtlich wird. Die Ähnlichkeit zwischen zwei Vertretern der Gruppe ist unterschiedlich und hängt von den Wechselbeziehungen ab, die sich im Laufe der Jahrhunderte entwickelt haben. Anhand von literarischen Texten als Quellenmaterial werden in dieser Arbeit lexikalische Gemeinsamkeiten und Unterschiede des Niedersorbischen und Polnischen analysiert. Die Studie konzentriert sich auf identische oder ähnliche Wortnotierungen und nimmt spezifische Unterschiede der Vergleichssprachen auf, wobei einige wiederkehrende Unterschiede berücksichtigt werden. In der Analyse wird eine Methode des lexikalischen Vergleichs angewendet, die auf der Grundlage von Levenshtein-Distanzen eine Bewertung von Ähnlichkeitsgraden auf lexikalischer Ebene ermöglicht. Die Textgrundlage ist das Werk „Der kleine Prinz“, das aus dem Französischen in beide Vergleichssprachen übersetzt wurde. Dabei werden sowohl Fälle identischer Notierung als auch vorab definierte Notationsunterschiede in Betracht gezogen.

# Working on Sets of Words as a Means of Determining the Scale of Similarity between Related Languages (Using the Example of Lower Sorbian and Polish)

Slavonic languages, compared with other Indo-European languages, exhibit a high level of similarity, characterized by, for example, common vocabularies. The similarity between two of their representatives varies and depends on the mutual relations that have occurred over the centuries. Using literary texts as source material, this paper analyzes lexical similarities and differences between Lower Sorbian and Polish. It focuses on identical and similar word spellings, and considers the specific features of both languages, while allowing for certain recurring differences. This research utilizes the method of lexical comparison based on the Levenshtein distances, which allow for an assessment of the degree of lexical similarity. The analysis was conducted using the text of “The Little Prince” as its basis, which has been translated from French into both the languages being compared. It considers both cases of identical spelling and allows for certain predeter-mined differences in spelling.

Serbski Sorbisches institut Institut

DOI
https://doi.org/10.59195/lp.2025.72-51

**Bartłomiej Szawulak, Tadeusz Lewaszkiewicz, Piotr Formanowicz**

# Operacje na zbiorach słów jako sposób na określenie skali podobieństwa pomiędzy pokrewnymi językami (na przykładzie języków dolnołużyckiego i polskiego)

## 1. Wprowadzenie

Każdy język naturalny jest nośnikiem informacji, ale też historii, zmian i oddziaływań, które wpływały na populację posługującą się nim. Badania jego charakterystyk, przeprowadzone w porównaniu z sąsiadującymi z nim językami, pozwalają na dogłębne zrozumienie procesów, którym dany język podlegał. Język polski i język dolnołużycki są dwoma blisko spokrewnionymi językami zachodniosłowiańskimi, reprezentującymi dwa skrajne przypadki pod względem liczby użytkowników. Podczas gdy język polski jest największym przedstawicielem swojej grupy, z dziesiątkami milionów użytkowników oraz bogatą literaturą, język dolnołużycki jest jednym z języków słowiańskich o najmniejszej liczbie użytkowników oraz o literaturze charakterystycznej dla społeczności mniejszościowej. Oba języki, mimo rozwijania się w izolacji od siebie, choć poddane podobnym wpływom, np. języka niemieckiego, wykazują wysokie podobieństwo leksykalne względem siebie. Dokładne badania relacji pomiędzy polskim a dolnołużyckim były ograniczone ze względu na ograniczoną liczbę specjalistów oraz dostępne dane pozwalające na przeprowadzenie odpowiednich analiz, np. w pracy (Blažek 2020) wskazywano na ograniczenia dostępnych słowników dla obu języków łużyckich. Dziś otwierają się nowe możliwości do badań języków słowiańskich mniejszościowych ze względu na wzrost dostępnych zasobów (np. korpusy dla języka kaszubskiego i dolnołużyckiego), jak również rosnące zainteresowanie tymi językami. Digitalizacja tych danych umożliwia zastosowanie szeregu metod informatycznych z celem ich wykorzystywania do operacji na tekstach. Pozwalają one na przeprowadzanie analiz, które dotychczas były poza zasięgiem językoznawców oraz na konfrontację z dotychczasowym konsensusem naukowym. W niniejszej pracy przeprowadzono analizę porównawczą słownictwa polskiego i dolnołużyckiego w celu określenia ich stopnia podobieństwa leksykalnego. Źródłem słownictwa są tłumaczenia *Małego Księcia*, które reprezentują współczesną formę języka i zawierają zróżnicowane leksykalnie słownictwo. Do oceny podobieństwa słownictwa wykorzystana została metoda oparta na odległości Levenshteina, polegająca na pomiarze liczby różnic pomiędzy wyrazami, wsparta analizą podobieństw fonetycznych.

## 2. Stan wiedzy

Lingwistyka porównawcza jest działem językoznawstwa badającym skalę różnic i podobieństw pomiędzy językami naturalnymi. W wyniku badań prowadzonych od XIX wieku opracowano klasyfikację języków słowiańskich, dzieląc je na podstawie pokrewieństwa na grupy (np. języki zachodniosłowiańskie) i podgrupy (np. grupa języków lechickich Blažek 2020). Analizy te w przeważającej większości były przeprowadzane na danych leksykograficznych, wspartych analizą kontekstu historycznego. W czasach współczesnych rozwój informatyki i postępująca cyfryzacja pozwalają na wykorzystanie istniejących już metod porównawczych na znacznie większą skalę niż jeszcze dekadę temu. Daje to możliwość ponownego przeprowadzenia analiz porównawczych uzupełnionych o

DOI: https://doi.org/10.59195/lp.2025.72-51

nowe zasoby i techniki. Jednocześnie możliwe jest skupienie się na analizie języków mniejszościowych, pomijanych często ze względu na brak odpowiednich danych. W niniejszej pracy przedstawiamy wyniki porównania leksykalnego z użyciem odległości Levenshteina (LEVENSHTEIN 1966) jako miary do oceny podobieństwa pomiędzy słowami występującymi w tekstach polskich i dolnołużyckich. Podobne badania z wykorzystaniem odległości Levenshteina były już przeprowadzone na językach reprezentujących tę samą gałąź rodziny języków indoeuropejskich, takich jak języki germańskie, romańskie i słowiańskie (HEERINGA et al. 2023). Badając kolejno dystans leksykalny, fonetyczny i syntaktyczny wykazano, że uzyskane za pomocą proponowanych metod wyniki pokrywają się z konsensusem naukowym opisującym relacje pokrewieństwa pomiędzy badanymi językami. (Jako przykład, dystans pomiędzy czeskim i słowackim wynosił ok. 8,1 %, pomiędzy polskim a czeskim 21,6 % a pomiędzy polskim i bułgarskim 39,2 %.) Innym ciekawym zastosowaniem odległości Levenshteina jest badanie wzajemnej zrozumiałości międzyjęzykowej (GOOSKENS/VAN HEUVEN 2017), gdzie zaobserwowano silną korelację między wynikami testów przeprowadzonych wśród rodzimych użytkowników badanych języków a wynikami opartymi na miarach odległości językowej (ważony wariant odległości Levenshteina). Niestety w obu wspomnianych badaniach porównaniu podlegali główni przedstawiciele tych rodzin językowych, a języki mniejszościowe (np. prowansalski dla rodziny języków romańskich czy język kaszubski dla języków słowiańskich) nie były w nim wzięte pod uwagę. Potwierdziły one jednak skuteczność opisanych metod, co pozwala zastosować je do badania relacji pomiędzy językiem polskim a dolnołużyckim.

## 3. Wybór tekstu

Literatura dolnołużycka jest typowym przykładem literatury stworzonej przez mniejszość etniczną zamieszkującą w większości tereny wiejskie. Obfituje ona we wspomnienia, listy oraz utwory o tematyce religijnej (tematyka ta dominowała w niej do 1945 roku; WROCŁAWSKA/WYSZOMIRSKA 1996). Tłumaczenia książek i tekstów z języka polskiego stanowią niewielki ułamek w porównaniu z językiem niemieckim czy czeskim. Przykładem może być opracowanie twórczości Adama Mickiewicza wraz z tłumaczeniem pięciu ksiąg *Pana Tadeusza* (NORBERG/KOSTA/MĚŠKANK 2013). Z czasów NRD dysponujemy też przekładami noweli i wierszy dolnołużyckich na język polski. Niestety, od początku lat 90-tych XX wieku obserwowany jest zanik tłumaczeń pomiędzy tymi językami. Wybór tekstu dostępnego w obu językach nie był więc kwestią prostą. Największe dzieło w piśmiennictwie dolnołużyckim stanowi ostatnie wydanie tłumaczenia Biblii ewangelickiej z roku 1868 (BIBLIJA 1868). Mimo że opublikowana ponad półtora wieku temu, miała ona i ma nadal duży wpływ na współczesny język literacki. Prezentuje jednak tekst religijny o charakterystycznym stylu i dużej liczbie archaizmów oraz słownictwa sakralnego, które odbiega od standardów języka nauczanego w szkołach i na kursach. Podobny zarzut może być skierowany pod adresem drugiej największej pozycji, czyli *Pana Tadeusza*, będącego XIX-wiecznym tekstem pisanym wierszem, który mimo wielkiego wpływu na polszczyznę odbiega od jej współczesnej formy. Analizując pozostałe teksty, będące nowelami, wierszami lub bajkami, wybór padł na książkę *Mały książę* autorstwa Antoine'a de Saint-Exupéry'ego.

W przeciwieństwie do stosunkowo licznej literatury dziecięcej występującej w obu językach a powstałej w czasach NRD, *Mały książę*, pod dolnołużyckim tytułem *Ten Mały*

*Princ,* został wydany dopiero w 2010 roku. Jest to bezpośrednie tłumaczenie z francuskiego oryginału, podobnie jak dwa tłumaczenia polskie analizowane w niniejszej pracy. Z tego powodu eliminowane jest zagrożenie interferencji językowej, która mogłaby zachodzić dla tłumaczenia z języka polskiego na dolnołużycki lub odwrotnie. Pozostałe pozycje literatury dziecięcej także nie reprezentują bezpośredniego tłumaczenia z polskiego lub dolnołużyckiego, a oryginalny tekst jest zapisany w języku górnołużyckim lub niemieckim. Przy szerszej analizie porównawczej należałoby sprawdzić, jaki jest wpływ języka oryginału na powstałe tłumaczenia. *Mały Książę* ma jednak znaczącą przewagę nad pozostałymi dostępnymi tekstami – jest jedną z najczęściej tłumaczonych na języki obce książek, w tym na języki mniejszości, takie jak języki łużyckie oraz na dialekty, jak np. dialekt wielkopolski. Tym samym daje możliwość dokonania porównań językowych z pozostałymi językami zachodniosłowiańskimi oraz rozpatrzenia uzyskanych wyników w kontekście innych języków słowiańskich. Jest to sytuacja rzadka w stosunku do pozostałych tekstów literatury dolnołużyckiej, które zazwyczaj występują jedynie w tłumaczeniach na niemiecki i górnołużycki. Wyjątki stanowią *Biblia* i *Pan Tadeusz*, których analizy porównawcze mogą być zrealizowane dopiero w przyszłości z opisanych powyżej powodów.

## 4. Porównanie leksykalne

Podejście prezentowane w niniejszej publikacji bazuje na porównaniu zbiorów słownictwa występującego w dwóch tłumaczeniach książki. Celem jest wyznaczenie zbioru słów o identycznym zapisie w obu badanych językach, pomniejszonego o wykryte przypadki fałszywego podobieństwa, a następnie określenie, jaki procent całego tekstu stanowią słowa wspólne dla obu języków. W ramach przeprowadzonych badań wyznaczono także dopasowania słów wykazujących pewne różnice w zapisie tych samych głosek za pomocą różnych liter w alfabetach polskim i dolnołużyckim. Przeprowadzone porównanie bazuje na identyczności zapisu, a nie identyczności wymowy porównywanych słów. Oba języki w wymowie głosek przedstawionych za pomocą tych samych liter są bardzo podobne, choć istnieje kilka różnic, których porównanie takie nie bierze pod uwagę (np. nagłos w słowach rozpoczynających się od *wo-*: dłuż. *wokno* – pol. *okno*).

Aby w pełni porównać słowa występujące w dwóch różnych językach, należałoby za pomocą słownika potwierdzić, że dopasowane słowa faktycznie mają to samo lub kontekstowo zbliżone znaczenie, lub że zastosowany został synonim istniejący w obu językach. Kolejnym krokiem może być usunięcie prefiksów i sufiksów (zbliżone działania do operacji stemmingu; Lovins 1968) w celu uzyskania rdzenia słowa niepodlegającego fleksji. Niestety, brak dostępu do cyfrowej wersji największego istniejącego słownika dolnołużyckiego i polskiego Leszczyńskiego (Leszczyński 2013) nie pozwala na automatyzację tego procesu. Dodatkowo sam słownik zawiera jedynie 8,5 tysiąca haseł. Dla porównania słownik dolnołużycko-niemiecki Starosty z 1999 roku zawiera 45 tysięcy haseł (Starosta 1999). W przyszłości, poza digitalizacją słownika Leszczyńskiego, potrzebne będzie jego rozszerzenie tak, aby uzyskać skuteczne narzędzie.

Z powodu opisanego powyżej braku odpowiednich słowników zdecydowaliśmy się na podejście heurystyczne, gdzie za pomocą reguł opisanych w sekcji 4.2.2 niniejszej pracy dopasowane zostały słowa uznane za podobne, a następnie z użyciem wiedzy eksperckiej podobieństwa te zostały potwierdzone.

Proces porównania opisany został w kolejnych podrozdziałach.

## 4.1 Porównanie

Dla każdej z wersji językowych tekstu, jaki podlega porównaniu, wyznaczany jest zbiór słów pozbawiony powtórzeń (duplikatów). Przez słowo rozumiemy ciąg liter ograniczony spacjami (pustymi znakami). Następnie wyznaczany jest zbiór słów występujących w obu tekstach w identycznym zapisie.

Tego rodzaju porównanie jest uzasadnione tylko wtedy, gdy dwa języki są ze sobą blisko spokrewnione i posiadają bardzo podobną formę zapisu. Jest to kwestia nierozerwalnie związana z historią tych języków oraz towarzyszącą jej polityką, która mogła dążyć do unifikacji zapisu (np. języki dolno- i górnołużyckie w latach 50. XX wieku, czeski i słowacki w okresie Czechosłowacji) lub zaznaczenia odrębności (np. ukraińska łacinka bazująca na zapisie czeskim, a nie sąsiednim polskim).

Drugim wariantem porównania jest wyznaczenie zbioru wspólnych słów po uwzględnieniu pewnych różnic w zapisie. Odnośnie języków polskiego i dolnołużyckiego wykazują one duże podobieństwo pomiędzy głoskami przypisanymi do odpowiadających sobie liter, przy istnieniu kilku różnic. Różnice te mogą polegać na tym, że różne litery służą do zapisu tej samej głoski (*ż* i *ž*), lub ta sama litera opisuje różne głoski w językach polskim i dolnołużyckim (*ó*).

Jednocześnie należy zaznaczyć, że porównaniu podlega tutaj sam zapis. Nie jest to porównanie po zastosowaniu transkrypcji do wspólnego zapisu fonetycznego. Taka analiza warta jest dokładniejszych badań w przyszłości.

Trzecim wariantem porównania jest znalezienie podobnych słów z uwzględnieniem charakterystycznych oboczności pomiędzy językami, opisanych w sekcji 4.2.2.

## 4.2 Wyniki porównania

Badaniu podlegały trzy warianty tekstu *Małego Księcia*. W wypadku języka dolnołużyckiego istnieje tylko jedno tłumaczenie Pětša Janaša, z roku 2010 (DE SAINT-EXUPÉRY 2010), wydane przez Edition Tintenfaß. Natomiast w języku polskim ukazały się aż dwadzieścia dwa tłumaczenia, z których w ramach tej publikacji rozważane są dwa: Jana Szwykowskiego z 1956 roku (DE SAINT-EXUPÉRY 2007) i Agaty Kozak z 2021 roku (DE SAINT-EXUPÉRY 2022).

Tekst Janaša posiada o 33 % więcej słów niż tłumaczenie Kozak i o 38 % więcej niż tłumaczenie Szwykowskiego (rozmiar polskich przekładów stanowi 75,20 % i 72,38 % rozmiaru dolnołużyckiego przekładu). Odpowiada za to składnia języka dolnołużyckiego, np. liczniejsze wystąpienia słowa *jo* (754 wystąpienia). Słowo to jest odpowiednikiem polskich słów *jest* (101 wystąpienia) oraz *tak* (50 wystąpień) i w języku dolnołużyckim (w pierwszym znaczeniu) pełni rolę czasownika posiłkowego, co wpływa na liczbę jego wystąpień. Jeśli natomiast pominiemy wielokrotne występowanie słów i zliczymy tylko pierwsze wystąpienia słów, pojawia się w zależności od tłumaczenia 6,67 %–7,93 % przewaga polskich słów nad dolnołużyckimi (Tabela 1). Jak już wcześniej wspomniano, porównania zostały przeprowadzone na zbiorach słów nie uwzględniających powtórzeń, wyznaczonych dla każdego z rozważanych tłumaczeń.

| Tłumacz | Liczba słów (uwzględniając powtórzenia) | Liczba unikalnych słów (bez powtórzeń) |
|---|---|---|
| P. Janaš | 14911 | 3254 |
| A. Kozak | 11213 | 3512 |
| J. Szwykowski | 10793 | 3471 |

Tabela 1: Stosunek liczby słów występujących w poszczególnych tłumaczeniach z podziałem na liczby słów z powtórzeniami (kolumna Liczba słów) i bez powtórzeń (Liczba unikalnych słów).

### 4.2.1 Różnice w alfabetach

Alfabet dolnołużycki składa się z 34 liter, z których 11 posiada znaki diakrytyczne. Z perspektywy języka polskiego i dolnołużyckiego może on zostać podzielony na pięć grup:

1) litery wraz z przypisanymi im głoskami, występujące w alfabecie dolnołużyckim – *ě, ŕ*;
2) litery występujące w alfabecie dolnołużyckim, odpowiadające głoskom występującym w języku polskim – *ž, š, č* (w przypadku litery *č* nie stanowi ona ekwiwalentu do polskiego *cz*, ale dźwięk zbliżony);
3) litera występująca w obu alfabetach, ale o innej przypisanej głosce – *ó*;
4) litery występujące w obu alfabetach i przedstawiające tę samą głoskę – *a, b, c, ć, e, f, g, i, j, k, l, ł, m, n, ń, o, p, r, s, ś, t, u, y, ź*;
5) litery o kilku przypisanych głoskach, zależnych od pozycji w słowie i jego pochodzenia – *w, h*.

W języku dolnołużyckim litera *w* jest wymawiana w większości przypadków jako głoska zbliżona do polskiego *ł*. Odstępstwem od tej reguły jest jej niema realizacja na początku wyrazu przed spółgłoskami oraz w większości przypadków na początku morfemu przed *o* lub *u*. Podobnie jest z literą *h*, która zasadniczo pozostaje niema we wszystkich pozycjach, ale opcjonalnie może być również wymawiana jako słabe *h* (KAULFÜRST/SZCZEPAŃSKA 2019). W Tabeli 2 przedstawiono przykłady dopasowanych słów.

| **Części mowy** | **Wspólne słowa w tłumaczeniach Janaša i Kozak** |
|---|---|
| Zaimki | ja, wam, tymi, ty, taka, takim, wy, je, nim, taki, tej, was, my, nich, nas, nikogo, nimi, nam, nami, *swój*, *mój*, *twój*, ma, jej, co, kim, kogo, komu, te, samo, sam, sami, to, nic, ten, tych, ta, tym, taki, nikomu, takich |
| Przyimki | we, z, dla, ze, nad, do, na, w, po, mimo, za, ku, nad |
| Przymiotniki | *żywy* (dłuż. żywy), stary, samotny, młodego, drugich, druga, małe, słabe, złote, mała, mały, małego, nowy, dobry, drugi, małych, bogaty, stara, *ważne* (dłuż. ważne), *starszy* (dłuż. staršy), słaby, małym, dobra, absolutny, małemu, nowego, staremu, złota |
| Przysłówki | tam, tak, mało, dalej, tu, *już* (dłuż. już), *bliżej* (dłuż. bliżej) |
| Spójniki | aby, a, co, ale, *aż* (dłuż. aż) |
| Czasowniki | mam, był, byli, była, dodał, zdało, było, *móc*, bywa, chowa, zapalił, pisał, *szli* (dłuż. šli), *służył* (dłuż. służył), płakał, dał, znał, mamy, daj, pytał, *słyszysz* (dłuż. słyšyš), pił, *waży* (dłuż. waży), zjawił, *masz* (dłuż. maš) |
| Rzeczowniki | głos, wina, planeta, numer, baobab, baobaby, planety, asteroida, dom, wulkan, wulkany, ptaka, droga, geograf, *góry*, geografa, oceany, *góra*, drogi, słowa, złoto, rano, boa, mil, rogi, nocy, planet, Mars, astronom, asteroid, planetach, |

| | |
|---|---|
| | palcami, barwy, wina, pomocy, głowy, banku, noc, karawana, jamy, *pszenica* (dłuż. pšenica), minutach, minuty, studni, baran, głodu, miliona, rady, winy |
| Wykrzyknik | aha, och, ach |
| Liczebnik | raz, dwa, sto |
| Partykuła | by, no |

Tabela 2: Wspólne słowa wyodrębnione w wyniku porównania tłumaczeń Janaša i Kozak z podziałem na części mowy. Kursywą zaznaczono słowa zawierające wystąpienia litery *ó*, która reprezentuje różne głoski w obu językach, jak i wystąpienia *ż* i *sz*, które zostały uznane za tożsame z *ž* i *š*.

W tym miejscu należy też zaznaczyć, że słowa przedstawione w Tabeli 2 nie reprezentują wszystkich sytuacji, gdy słowa można uznać za podobne do siebie. W przypadku słowa *tej*, które w dolnołużyckim może odpowiadać formie podwójnej, sprawdzono, czy w tekstach występuje ono w tej samej roli (dłuż. *Ty sy kradu słaby tud na tej Zemi z granita.* — pol. … *i znalazłeś się na tej Ziemi zbudowanej z granitu.*). W języku dolnołużyckim występuje duża liczba słów posiadających identyczne lub zbliżone odpowiedniki w języku polskim, które obecnie określane są jako archaizmy. Część z nich to zapożyczenia z języka niemieckiego, które mimo że występowały w przeszłości w języku polskim, to we współczesnej polszczyźnie zostały wyparte przez inne słowa. Przykładem jest dłuż. *štymujo* – pol. *sztymuje* – niem. *stimmen*. Aby uchwycić takie przypadki w tekstach, konieczna jest baza powiązań stworzona na podstawie synonimów, co obecnie nie jest możliwe dla tej pary porównywanych języków. Powodem jest brak odpowiednich słowników synonimów pozwalających taką bazę stworzyć.

### 4.2.2 Akceptowalne różnice

Języki te, jako blisko spokrewnione, wykazują systematyczne podobieństwa i różnice zachodzące między nimi. Przykładem niech będzie bezokolicznik dłuż. *pisaś* i pol. *pisać*, gdzie *ś* odpowiada polskiemu *ć* (STIEBER 1965). Podobne procesy są obecne w dialektach języka polskiego w stosunku do polskiego języka literackiego (np. mazurzenie, cakanie) i są efektem naturalnych zmian, obserwowanych w continuum językowym.

Bazując na wiedzy o obu językach, opracowano zbiór par liter obrazujących powtarzalną różnicę fonetyczną. Dodatkowo przeprowadzono analizę porównawczą z wykorzystaniem odległości Levenshteina (LEVENSHTEIN 1966), którą wykonano na zbiorach nie powtarzających się słów z obu języków. Przeanalizowano pary słów dla których dystans wynosił kolejno 1, 2, 3, 4 (para słów *dom* - *doma* różni się jedną literą, stąd dystans Levenshteina dla niej wynosi 1, natomiast dla pary *pisać* - *pisaś* dystans wynosi 2, ponieważ wymaga dwóch operacji w celu przekształcenia jednego słowa w drugie tj. usunięcia litery *ś* i wstawienia litery *ć*). Zaobserwowano, że nawet dla niewielkiego dystansu (1 i 2) liczba błędnych dopasowań była duża (ponad 50 %). Jednocześnie potwierdzono wyznaczone wcześniej relacje pomiędzy literami oraz uzupełniono je o nowo zaobserwowane. Z tego powodu podczas wyznaczania nowego zbioru par dopuszczalny dystans został ustalony na 0, a w przypadku występowania liter reprezentujących powtarzalną różnicę fonetyczną były one traktowane jak identyczne litery (przykładem jest para dłuż. *lěpjej* – pl. *lepiej*, gdzie *ě* i *e*, oraz *j* i *i* są traktowane jako pary liter identycznych). Przy zastosowaniu takich ograniczeń liczba błędnych dopasowań spadła do 44 przypadków reprezentujących około 19 % wszystkich znalezionych dopasowań, gdzie 12 % stanowiły dopasowania słów o takim samym znaczeniu i zapisie, ale innej odmianie, a pozostałe 6 % reprezentowały niepoprawne dopasowania. Ten poziom błędnych dopasowań został uznany za akceptowalny. Warto wspomnieć, że wykryte zostały również nieliczne błędy

drukarskie w dolnołużyckim tłumaczeniu, jak np *moj, won, wo*, które przez algorytm zostały błędnie zakwalifikowane jako osobne przypadki i dopasowane w pary z polskimi słowami *mój*, *on*, *o*.

| Relacja | Częstość | Unikalne pary | Przykład |
|---|---|---|---|
| ó → o | 46 | 46 | wón → _on<br>mójej → mojej |
| j → i | 41 | 36 | njej → niej |
| o → e | 29 | 29 | togo → tego |
| u → ę | 23 | 22 | wuže → węże |
| o → ó | 23 | 23 | rowno → równo |
| c → cz | 22 | 22 | coło → czoło |
| ś → ć | 20 | 20 | pytaś → pytać |
| s → ś | 15 | 15 | šesć → sześć |
| i → y | 14 | 13 | pśi → przy |
| ś → rz | 11 | 11 | pśi → przy |
| u → a | 8 | 7 | kuždе → każde |
| y → i | 7 | 7 | wóny → _oni |
| y → o | 5 | 5 | słyńca → słońca |
| ě → a | 5 | 5 | lět → lat |
| ě → e | 5 | 5 | lěpjej → lepiej |
| a → ó | 4 | 4 | kral → król |
| r → rz | 3 | 3 | grib → grzyb |
| z → ź | 2 | 2 | grozne → groźne |
| ji → i | 2 | 2 | jich → ich |
| i → u | 2 | 2 | sni → snu |
| z → ż | 1 | 1 | prozne → próżne |
| ć → cz | 1 | 1 | płašća → płaszcza |
| ć → c | 1 | 1 | złośći → złości |

Tabela 3: Relacje przedstawiające akceptowalne różnice wraz z częstotliwością ich występowania, liczbą par, w których wystąpiły, oraz przykładową parą.

W przypadku czterech rozważanych powiązań liter i dwuznaków dolnołużyckich z polskimi: *z → ź*, *si → s*, *ci → c*, *ci → ć*, w badanym tekście nie znaleziono żadnych słów zawierających je, dla których dystans wynosiłby 0. Takie relacje nadal mogły zachodzić pomiędzy badanymi zbiorami słów, jednak w takich przypadkach wystąpić musiały dodatkowe różnice w zapisie.

Warty uwagi jest przypadek pary *lěc → lecz*, która spełnia relację zawartą w Tabeli 3, jednak nie jest klasyfikowana jako para słów o tym samym znaczeniu. Etymologia wyrazów jest taka sama, jednak dziś w obu językach mają one inne znaczenia. Dlatego we współczesnych tekstach nie odpowiadają one sobie i nie są klasyfikowane jako para odpowiadających sobie słów w ramach przeprowadzonych porównań.

| | Akceptowalne dopasowania dolnołużycki – polski | Liczba |
|---|---|---|
| Janaš – Kozak | wón - on, togo - tego, šesć - sześć, lět - lat, wuže - węże, swójo - swoje, mója - moja, mójej - mojej, njej - niej, wuža - węża, karieru - karierę, pytaś - pytać, cas - czas, pśi - przy, pśez - przez, mójo - moje, rowno - równo, byś - być, tomu - temu, jogo - jego, człowjeka - człowieka, mnjo - mnie, daloko - daleko, wjele - wiele, mnje - mnie, nakresliś - nakreślić, bog - bóg, wócy - oczy, drogu | 177 |

| | | |
|---|---|---|
| | - drogę, njomu - niemu, ważnjejše - ważniejsze, domyslił - domyślił, mójogo - mojego, wóna - ona, póznał - poznał, móje - moje, kuždе - każde, psikład - przykład, lěta - lata, lětaś - latać, widaś - widać, głowu - głowę, mózo - może, mysl - myśl, wokoło - około, smjaś - śmiać, wóno - ono, dajo - daje, diktator - dyktator, swójomu - swojemu, licby - liczby, woknach - oknach, gołubjami - gołębiami, musyš - musisz, frankow - franków, wóni - oni, smjał - śmiał, kuždy - każdy, budu - będę, lěpjej - lepiej, pódobny - podobny, dnju - dniu, twója - twoja, mysli - myśli, baobabow - baobabów, samogo - samego, katastrofu - katastrofę, tśi - trzy, riziko - ryzyko, kuždego - każdego, pśed - przed, twójo - twoje, słyńca - słońca, słyńco - słońce, minutu - minutę, smjerś - śmierć, spódoba - spodoba, psigótowanja - przygotowania, comu - czemu, dosć - dość, wódy - wody, licyś - liczyć, ruce - ręce, cłowjek - człowiek, grib - grzyb, milionow - milionów, cogo - czego, njogo - niego, ruki - ręki, twójej - twojej, twóje - twoje, pokazaś - pokazać, casu - czasu, wó - o, pód - pod, słuchaś - słuchać, nje - nie, swóje - swoje, płaco - płacze, kral - król, tronje - tronie, kralom - królem, zacerwjenił - zaczerwienił, daś - dać, płašća - płaszcza, cym - czym, krala - króla, napisaś - napisać, ruku - ręku, połnych - pełnych, pomoc - pomóc, zapaliś - zapalić, kónca - końca, mógu - mogę, nicogo - niczego, latarnju - latarnię, zgasył - zgasił, zgasyś - zgasić, coło - czoło, spaś - spać, dny - dni, słyńcom - słońcem, daloka - daleka, rěki - rzeki, licy - liczy, zdajo - zdaje, moralnosć - moralność, kamjenje - kamienie, wopisaś - opisać, rědko - rzadko, kralow - królów, geografow - geografów, poł - pół, tśista - trzysta, latarnje - latarnie, africe - afryce, połnocnej - północnej, wóne - one, psišło - przyszło, góru - górę, rože - róże, rožu - różę, stśělby - strzelby, swójej - swojej, muzika - muzyka, pšenicne - pszeniczne, póla - pola, póznaś - poznać, gótowe - gotowe, psigótowaś - przygotować, dnja - dnia, płakaś - płakać, rožami - różami, roža - róża, ważnjejša - ważniejsza, studnju - studnię, słowka - słówka, wóda - woda, studnja - studnia, wobraz - obraz, studnje - studnie, rožy - róży, zasmjał - zaśmiał, muri - muru, domysliś - domyślić, žołty - żółty, kamjenjami - kamieniami, wužom - wężem, pójdu - pójdę, smjejoš - śmiejesz, grozne - groźne, wokno - okno, zlě - źle, chóry - chory, minuło - minęło, jich - ich, jim - im | |
| Janaš – Szwykowski | wón - on, togo - tego, šesć - sześć, lět - lat, wuže - węże, móju - moją, mója - moja, mójej - mojej, njej - niej, wuža - węża, karieru - karierę, pytaś - pytać, cas - czas, pśi - przy, pśez - przez, mójo - moje, rowno - równo, byś - być, tomu - temu, jogo - jego, cłowjeka - człowieka, mnjo - mnie, wódu - wodę, daloko - daleko, wjele - wiele, mnje - mnie, wócyma - oczyma, wócy - oczy, pśede - przede, drogu - drogę, ważnjejše - ważniejsze, domyslił - domyślił, wóna - ona, póznał - poznał, móje - moje, psikład - przykład, lěta - lata, cynjenja - czynienia, głowu - głowę, mózo - może, mysl - myśl, wokoło - około, smjaś - śmiać, wóno - one, dajo - daje, diktator - dyktator, swójomu - swojemu, woknach - oknach, gołubjami - gołębiami, musyš - musisz, wóni - oni, kuždy - każdy, budu - będę, lěpjej - lepiej, pódobny - podobny, dnju - dnia, twója - twoja, mysli - myśli, baobabow - baobabów, móžoš - możesz, tśi - trzy, riziko - ryzyko, kuždego - każdego, pśed - przed, górne - godne, twójo - twoje, słyńca - słońca, słyńco - słońce, minutu - minutę, smjerś - śmierć, spódoba - spodoba, psigótowanja - przygotowania, wódy - wody, złosći - złości, ruce - ręce, carnymi - czarnymi, cłowjek - człowiek, grib - grzyb, woblicu - obliczu, milionow - milionów, cogo - czego, njogo - niego, twójej - twojej, twóje - twoje, pokazaś - pokazać, wó - o, pód - pod, słuchaś - słuchać, swóje - swoje, płaco - płacze, kral - król, purpuru - purpurę, tronje - tronie, znaś - znać, zacerwjenił - zaczerwienił, daś - dać, płašća - płaszcza, cym - czym, krala - króla, napisaś - napisać, ruku - ręku, połnych - pełnych, pomoc - pomóc, zapaliś - zapalić, mógu - mogę, nicogo - niczego, latarnja - latarnia, latarnju - latarnię, zgasył - zgasił, zgasyś - zgasić, coło - czoło, spaś - spać, dny - dni, słyńcom - | 166 |

| | | |
|---|---|---|
| | słońcem, daloka - daleka, rěki - rzeki, licy - liczy, zdajo - zdaje, moralnosć - moralność, kamjenje - kamienie, wopisaś - opisać, rědko - rzadko, won - on, moj - mój, kralow - królów, geografow - geografów, poł - pół, tśista - trzysta, kontinentach - kontynentach, połnocnej - północnej, wóne - one, góru - górę, rože - róże, rožu - różę, stśělby - strzelby, swójej - swojej, muzika - muzyka, dnja - dnia, póznaś - poznać, gótowe - gotowe, dnja - dnia, płakaś - płakać, prozne - próżne, rožami - różami, roža - róża, rozjasnił - rozjaśnił, studnju - studnię, wóda - woda, studnja - studnia, wobraz - obraz, studnje - studnie, rožy - róży, zasmjał - zaśmiał, głowki - główki, muri - muru, wužom - wężem, pójdu - pójdę, smjejoš - śmiejesz, grozne - groźne, wokno - okno, złě - źle, chóry - chory, minuło - minęło, jich - ich, jim - im, njebudu - nie będę | |
| Pary występujące tylko w Janaš – Kozak | swójo - swoje, nakresliś - nakreślić, bog - bóg, njomu - niemu, mójogo - mojego, kuždе - każde, lětaś - latać, widaś - widać, wóno - ono, licby - liczby, frankow - franków, smjał - śmiał, dnju - dniu, samogo - samego, katastrofu - katastrofę, comu - czemu, dosć - dość, licyś - liczyć, ruki - ręki, kralom - królem, kónca - końca, latarnje - latarnie, africe - afryce, pśišło - przyszło, pšenicne - pszeniczne, póla - pola, pśigótowaś - przygotować, ważnjejša - ważniejsza, słowka - słówka, domysliś - domyślić, žołty - żółty, kamjenjami - kamieniami, casu - czasu, nje - nie | 34 |
| Pary występujące tylko w Janaš – Szwykowski | móju - moją, wódu - wodę, wócyma - oczyma, pśede - przede, cynjenja - czynienia, dnju - dnia, móžoš - możesz, gódne - godne, złości - złości, carnymi - czarnymi, woblicu - obliczu, purpuru - purpurę, znaś - znać, latarnja - latarnia, moj - mój, kontinentach - kontynentach, rozjasnił - rozjaśnił, głowki - główki | 18 |

Tabela 4: Pary słów dolnołużyckich i polskich o akceptowalnych różnicach, opisanych w sekcji 4.2.2.

Rezultatem opisanego porównania są dwa zbiory o licznościach 166 i 177 par słów o nieidentycznym zapisie, ale zawierającym charakterystyczne podobieństwa pomiędzy językiem polskim a dolnołużyckim. W Tabeli 4 przedstawiono wyznaczone pary z podziałem na tłumaczenia.

Pary dopasowanych słów o identycznym lub zbliżonym zapisie i akceptowalnych różnicach zostały sprawdzone słownikowo pod względem zgodności znaczenia. Stwierdzono kilkadziesiąt przypadków niepoprawnego dopasowania, np. *pytaś – pytać, jědu – jadu* (inne znaczenie słów), *południu – południu, wót – ot, skóro – skoro* (inne znaczenie kontekstowe), *wódu – woda, kralowy – królowi* (słowa będące w innej odmianie, innym przypadku lub będącej inną częścią mowy). Sytuacje takie nie zostały uwzględnione w opisanym powyżej porównaniu.

### 4.3.4 Podsumowanie

Wyniki opisanych powyżej analiz słownictwa z podziałem na pary porównywanych tłumaczeń zostały przedstawione w Tabeli 5. Wiersze tej tabeli odpowiadają kolejno rozważanym wariantom. Pierwszy dotyczy przypadków, gdy zapis słów jest identyczny (*idealne dopasowanie*). Drugi wiersz odpowiada przypadkom, w których zapis słów został ujednolicony ze względu na różnice w alfabetach (*ujednolicony zapis*). W tych przypadkach wyszczególniono trzy dodatkowe sytuacje: pierwsza, gdy różne litery odpowiadające tej samej głosce zostały zunifikowane: *š=sz*, *ž=ż*, *č=cz*, druga dla litery *ó*, odpowiadającej różnym głoskom w języku polskim i dolnołużyckim (podrozdział 4.2.1) i trzecia będąca połączeniem obu poprzednich. Kolejny wiersz (*Akceptowalne różnice*) zawiera

liczby stwierdzonych par słów, w których istnieją akceptowalne różnice ustalone w podrozdziale 4.2.2. Ostatni wiersz zawiera sumę wartości uzyskanych dla akceptowalnych różnic i przypadku 3) z ujednoliconego zapisu.

| | Janaš – Szwykowski | Janaš – Kozak | Szwykowski – Kozak |
|---|---|---|---|
| idealne dopasowanie | 158 (4,55 %) | 161 (4,58 %) | 2057 (58.5%) |
| ujednolicony zapis<br>1) š=sz, ž=ż, č=cz<br>2) pl. ó≠ dłuż. ó<br>3) przypadki 1) + 2) | <br>170 (4,90 %)<br>153 (4,41 %)<br>165 (4,75 %) | <br>173 (4,93 %)<br>155 (4,41 %)<br>167 (4,76 %) | |
| akceptowalne różnice | 166 (4,78 %) | 177 (5,04 %) | |
| ujednolicony zapis 3)<br>+ akceptowalne różnice | 331 (9,54 %) | 344 (9,79 %) | |

Tabela 5: Rezultaty porównania zbiorów słów bez powtórzeń dla par porównywanych tłumaczeń przedstawiono w trzech odpowiadających im kolumnach. Dwie z nich przedstawiają porównanie pomiędzy tekstem dolnołużyckim a polskim, podczas gdy w trzeciej zamieszczono wynik porównania dwóch polskich tłumaczeń. Wynik ten jest używany jako punkt referencyjny do interpretacji pozostałych rezultatów.

Analizując wyniki zamieszczone w Tabeli 5 można zauważyć, że wartości uzyskane dla porównań różnych tłumaczeń na język polski z tłumaczeniem na język dolnołużycki są zbliżone rozmiarem. Warto zaznaczyć, że zbiory par o akceptowalnych różnicach stwierdzone dla różnych tłumaczeń współdzielą ze sobą 80 % i 90 % par (kolejno dla Kozak i Szwykowskiego). W przypadku ujednoliconego zapisu, różnice w zapisie w alfabetach takie jak *š* i *sz* mają bardzo ograniczony wpływ na liczbę dopasowań.

Zbiór zawierający pary o identycznym zapisie oraz pary o akceptowalnych różnicach (np. *pisaś* = *pisać*) ma wielkość 331 w przypadku porównania z tłumaczeniem Szwykowskiego i 344 z tłumaczeniem Kozak. Odpowiada to 9,54 % i 9,79 % wszystkich wystąpień słów bez powtórzeń. Poziom ten należy uznać za znaczący. Zwrócić także należy uwagę na fakt, że pomiędzy oboma tłumaczeniami na język polski identyczność słownictwa wynosi 58,5 %. To ta wartość, a nie 100 %, które odpowiadałyby dwóm identycznym tekstom, powinna być użyta jako punkt odniesienia dla wyników uzyskanych przy porównaniach międzyjęzykowych. W przypadku analizy dopuszczającej wielokrotne wystąpienia słów współczynnik współdzielonego słownictwa będzie wyższy ze względu na to, że dopasowane słowa reprezentują podstawowe słownictwo, dla którego istnieje większe prawdopodobieństwo wielokrotnego wystąpienia w tekście.

## 5. Wnioski

Na podstawie przeprowadzonych badań analogicznych tekstów w różnych językach można stwierdzić, że podobieństwo leksykalne pomiędzy językiem polskim a językiem dolnołużyckim ma pewien stały poziom i to pomimo ponad pół wieku, jakie dzieli powstanie obu polskich tłumaczeń. Nie jest możliwe stwierdzenie, jakie zmiany zaszły w tym okresie w języku dolnołużyckim, jako że istnieje tylko jedno wydanie książki *Ten Mały Princ* z początku XXI wieku. Natomiast można wnioskować, że zmiany, jakie zaszły w języku polskim, nie miały znacznego wpływu na część słownictwa współdzielonego z językiem dolnołużyckim. Analizując słowa uznane za podobne (Tabele 2–4)

można stwierdzić, że reprezentują one wszystkie części mowy oraz szerokie spektrum leksyki – od terminów dotyczących przyrody po techniczne internacjonalizmy.

Przeprowadzone w niniejszej pracy porównania stanowią jedynie pierwszy krok w badaniu podobieństwa występującego pomiędzy językami polskim i dolnołużyckim. Kolejnym krokiem powinno być przeprowadzenie porównania tłumaczeń na inne języki zachodniosłowiańskie, zwłaszcza górnołużycki, z którym dolnołużycki jest ściśle spokrewniony, czeski, który wpływał bezpośrednio i pośrednio na język dolnołużycki przez siostrzany język górnołużycki oraz kaszubski, będący przedstawicielem dialektów pomorskich, cechujący się ponadto znacznymi wpływami niemieckimi. Pozwoliłoby to na odpowiednie określenie skali podobieństwa poprzez przeprowadzenie analogicznego porównania tekstów zapisanych w innych językach zachodniosłowiańskich. Potwierdzenie lub zakwestionowanie istnienia lub skali opisywanych relacji jest istotne z perspektywy języków słowiańskich o małej liczbie użytkowników. Drugim krokiem powinno być wykorzystanie istnienia licznych tłumaczeń *Małego Księcia* na język polski i dokonanie szerszego ich porównania. Pozwoli to na zmniejszenie wpływu języka i stylu tłumacza na końcową ocenę średniego poziomu podobieństwa.

Jako kolejny krok sugerujemy wprowadzenie technik reprezentacji grafowej, np. dostosowanie reprezentacji Universal Dependencies (DE MARNEFFE et al. 2021) do języków łużyckich i przeprowadzenie z jej zastosowaniem porównań relacji składniowych pomiędzy językami słowiańskimi.

Mamy nadzieję, że niniejsza praca zainteresuje środowisko językoznawców (w tym sorabistów) oraz będzie stanowić punkt wyjścia do dalszych badań.

## Bibliografia

BIBLIJA 1868: Biblija abo to zełe Sswéte Pißmo Starego a Nowego Testamenta, do ßerbskeje rězy pschestawjone. Wot nowotki pilńe pscheglėdane a pscheporėžane. Halle. https://niedersorbisch.de/biblija/ [28.11.2025].

BLAŽEK, Vaclav 2020: Classification of Slavic Languages: Evolution of Developmental Models, w: Slavia Occidentalis 77/1, s 33–64.

DE MARNEFFE, Marie-Catherine; MANNING, Christopher D.; NIVRE, Joakim; ZEMAN, Daniel 2021: Universal Dependencies, w: Computational Linguistics 47/2, s. 255–308.

DE SAINT-EXUPÉRY, Antoine 2007: Mały Książę; tłum. Jan SZWYKOWSKI. Warszawa.

DE SAINT-EXUPÉRY, Antoine 2010: Ten Mały Princ; tłum. Pětš JANAŠ. Neckarsteinach.

DE SAINT-EXUPÉRY, Antoine 2022: Mały Książę; tłum. Agata KOZAK. Warszawa.

GOOSKENS, Charlotte; VAN HEUVEN, Vincent J. 2017: Measuring Cross-Linguistic Intelligibility in the Germanic, Romance and Slavic Language Groups, w: Speech Communication 89, s 25–36.

HEERINGA, Wilbert; GOOSKENS, Charlotte, VAN HEUVEN, Vincent J. 2023: Comparing Germanic, Romance and Slavic: Relationships among Linguistic Distances, w: Lingua 287, 103512.

KAULFÜRST, Fabian; SZCZEPAŃSKA, Joanna 2019: Dolnoserbske wugronjenje – Niedersorbische Aussprache. http://www.dolnoserbski.de/wugronjenje/ [25.03.2024].

LESZCZYŃSKI, Rafał 2013: Podręczny słownik polsko-dolnołużycki/dolnołużycko-polski. Budyšyn.

LEVENSHTEIN, Vladimir I. 1966: Binary Codes Capable of Correcting Deletions, Insertions, and Reversals, w: Soviet Physics Doklady, 10 (8), s. 707-710.

LOVINS, Julie Beth 1968: Development of a Stemming Algorithm, w: Mechanical Translation and Computational Linguistics 11/1–2, s 22–31.

NORBERG, Madlena; KOSTA, Peter; MĚŠKANK, Alfred 2013: Adam Mickiewicz. Basni w serbskich pśełožkach (= Podstupimske Pśinoski k Sorabistice / Potsdamer Beiträge zur Sorabistik; 10). Potsdam.

STAROSTA, Manfred 1999: Dolnoserbsko-nimski słownik. Niedersorbisch-deutsches Wörterbuch. Budyšyn/Bautzen.

STIEBER, Zdzisław 1965: Zarys dialektologii języków zachodnio-słowiańskich. Warszawa.

WROCŁAWSKA, Elżbieta; WYSZOMIRSKA, Agnieszka 1996: Z badań nad historią języków łużyckich (prace z drugiej połowy XX wieku), w: FASSKE, Helmut; WROCŁAWSKA, Elżbieta (red.), Z historii języków łużyckich (= Język na pograniczach 18). Warszawa: s. 13–52.